HEALTHCARE COST AND UTILIZATION PROJECT — HCUP
A FEDERAL-STATE-INDUSTRY PARTNERSHIP IN HEALTH DATA
Sponsored by the Agency for Healthcare Research and Quality

Technical Supplements for

The HCUP Nationwide Inpatient Sample (NIS), Release 6, 1997

Agency for Healthcare Research and Quality
2101 East Jefferson Street, Suite 605
Rockville, Maryland 20852

phone: (301) 594-3075
fax: (301) 594-2166
Internet: hcup@ahrq.gov

NIS Data and Documentation Distributed by:
National Technical Information Service
5285 Port Royal Road
Springfield, VA 22161

phone: 800-553-6847 or 703-605-6000

Table of Contents

TECHNICAL SUPPLEMENT 10:

TECHNICAL SUPPLEMENT 11:

TECHNICAL SUPPLEMENT 12:

Table of Tables

TECHNICAL SUPPLEMENT 1: DATA USE AGREEMENT FOR THE HCUP NATIONWIDE INPATIENT SAMPLE

Before using the HCUP Nationwide Inpatient Sample (NIS), all users must sign a copy of the Data Use Agreement that follows.

The Data Use Agreement stipulates that individuals who obtain the HCUP NIS may not release any part of the NIS to someone in another organization, except with the approval of AHCPR.

Individuals who obtain NIS data may share the data with others in their organization, but all users must sign the Data Use Agreement. Individuals who wish to share NIS data have the responsibility to:

- copy the Data Use Agreement,
- ensure that all users in their organization sign the agreement,
- keep copies of the signed agreements, and
- make the signed agreements available to AHCPR upon request.

For more details, please see the Data Use Agreement.

Data Use Agreement for HCUP Nationwide Inpatient Sample for Release

Under section 903(c) of the Public Health Service Act (42 U.S.C. 299a-1), data collected by the Agency for Health Care Policy and Research (AHCPR) may be used only for the purpose for which they were collected. Data supplied to AHCPR under the auspices of HCUP were provided by the data sources only for the purpose of research.

Person identifiers--Any effort to determine the identity of any person or to use the information for any purpose other than for analysis and aggregate statistical reporting would violate the AHCPR statute (above) and the conditions of this data use agreement. Furthermore, under the statute, no information may be published or released in any way if a person, who supplied the information or who can be identified by the information, has not consented to its release. AHCPR omits from the data set all direct personal identifiers, as well as characteristics that might lead to identifications of persons. It may be possible in rare instances, through complex analysis and with outside information, to ascertain from the data sets the identity of particular persons. Considerable harm could ensue if this were done. By virtue of this agreement, the undersigned agrees that such attempts will not be made and that in any event such information would never be released or published.

Establishment identifiers--Section 903(c) of the Public Health Service Act (42 U.S.C. 299a-1) also restricts the use of any information that allows the identification of establishments to the purpose for which the information was collected. Permission was obtained from data sources (state data organizations, hospital associations, and data consortia) to use the identification of hospitals (when such identification appears in the data sets) for the purpose of conducting research only. Such research purpose includes linking institutional information from outside data sets for analysis and aggregate statistical reporting. Such purpose does *not* include the use of information in the data sets concerning individual establishments for commercial or competitive purposes involving those individual establishments, or to determine the rights, benefits, or privileges of establishments. No establishments can be identified directly or by inference in disseminated material. Users of the data shall not contact establishments for the purpose of verifying information supplied in the HCUP database. Any questions about the data must be referred to AHCPR only.

The undersigned gives the following assurances with respect to the AHCPR data sets.

- I will not use nor permit others to use the data in these sets in any way except for research and aggregate statistical reporting;

- I will require others in the organization (specified below) who use the data to sign this agreement and will keep those signed agreements and make them available to AHCPR upon request;

- I will not release nor permit others to release any information that identifies persons, directly or indirectly.

- I will not release nor permit others to release the data sets or any part of them to any person who is not a member of the organization (specified below), except with the approval of AHCPR;

- I will not attempt to link nor permit others to attempt to link the hospital stay records of persons in this data set with personally identifiable records from any other source;

Data Use Agreement for HCUP Nationwide Inpatient Sample for Release (Continued)

- I will not attempt to use nor permit others to use the data sets to learn the identity of any person included in any set;

- I will not use nor permit others to use the data concerning individual establishments (1) for commercial or competitive purposes involving those individual establishments, (2) to determine the rights, benefits, or privileges of individual establishments nor (3) to report, through any medium, data that could identify, directly or by inference, individual establishments;

- When the identities of establishments are not provided on the data sets, I will not attempt to use nor permit others to use the data sets to learn the identity of any establishment in the data sets;

- I will not contact nor permit others to contact establishments or persons in the data sets to question, verify, or discuss data in the HCUP database;

- I will make no statement nor permit others to make statements indicating or suggesting that interpretations drawn are those of data sources or AHCPR; and

- I will acknowledge in all reports based on these data that the source of the data is the "Healthcare Cost and Utilization Project (HCUP), Agency for Health Care Policy and Research."

My signature indicates my agreement to comply with the above-stated requirements with the knowledge that deliberately making a false statement in any matter within the jurisdiction of any department or agency of the Federal Government violates 18 U.S.C. 1001 and is punishable by a fine of up to $10,000 or up to 5 years in prison. Violators of this agreement may also be subject to penalties from state statutes that apply to these data for particular states.

Signed: ________________________ Date: ____________

Print or Type Name: ________________________

Title: ________________________

Organization: ________________________

Address: ________________________

City: ________________ State: ________ Zip code: ________

Phone Number: ________________________

Note: The person who signs this data use agreement must be the person to whom the data product is shipped.

Rev. 1/22/97

TECHNICAL SUPPLEMENT 2:
QUALITY CONTROL IN HCUP DATA PROCESSING

This Technical Supplement describes the processes used to ensure the quality of HCUP Nationwide Inpatient Sample (NIS) data. It describes the quality review guidelines employed in reviewing data for each variable in the NIS, including the edit checks performed to assess the internal consistency of information on each record.

QUALITY REVIEW GUIDELINES

Table 1 (on page 15) summarizes the HCUP quality review guidelines. These guidelines were developed for use with both the NIS and the State Inpatient Database (SID). As a result, this table includes numerous variables that are not part of the NIS, but which may be obtained from some SID data sources. For example, this Technical Supplement refers to DCCHPR1-DCCHPR30. The NIS contains only DCCHPR1, while some states in the SID may include DCCHPR2 up to DCCHPR30, depending on how many diagnoses are provided by that data source. For more information about the SID, see *General Information About HCUP*.

These guidelines apply to the following summary statistics generated on large inpatient databases:

- number of missing values,
- minimum,
- maximum,
- mean, and
- frequency distributions.

The minimum and maximum values specified above are HCUP limits that may not occur for each data source.

Table 1. HCUP Variables – Quality Review Guidelines

Variable Name	Description	Guidelines
ADATE	Admission date.	Monthly frequencies should not fluctuate greatly. There may be some seasonal fluctuations – e.g., fewer admissions in the summer months.

ADAYWK	Admission day of week, Sunday to Saturday.	Missing as often as admission date, if calculated.
		Minimum = 1
		Maximum = 7

ADRG	All-patient refined DRG.	None.
ADRGSEV	All-patient refined DRG severity level.	None.

Table 1. HCUP Variables – Quality Review Guidelines

Variable Name	Description	Guidelines
AGE	Age in years at admission.	Few missing values. Expected mean = 40 Minimum = 0 Maximum = 124 If the mean is less than 40, look for a high percentage of births. If the mean is greater than 40, look for a low percentage of births, or a high percentage of Medicare patients. The distribution of age should be faintly trimodal with few values over 90. Since 10-13% of all admissions are births, approximately that many discharges will indicate age 0. The next swell in the frequency will appear in the childbearing years, 14-43. The third rise in the frequency will appear in the 49-72 age range.
AGEDAY	Age in days (coded only when the age in years is less than 1).	Many missing values. Should be coded for less than 20% of the observations. Expected mean = 15-20 Minimum = 0 Maximum = 364 Records with AGEDAY = 0 (newborns) will account for approximately 10-13% of all records.

AMONTH	Admission month,
	January to December.
		Missing less than or equal to the number of records missing admission dates.
		Minimum = 1
		Maximum = 12
		Monthly frequencies should not fluctuate greatly.

ASCHED	Scheduled vs.
	unscheduled admission.
		May have many missing values.
		Minimum = 0
		Maximum = 1

ASOURCE	Admission source includes emergency department, another hospital, other health facility, court/law enforcement, and routine.	Can have numerous missing values.
		Minimum = 1
		Maximum = 5
		Most records will be routine, birth, and other sources. The emergency department is the next most frequent source.

Table 1. HCUP Variables – Quality Review Guidelines

Variable Name	Description	Guidelines
ATYPE	Admission type includes emergency, urgent, elective, newborn, and delivery.	Can have numerous missing values. Minimum = 1 Maximum = 6 In most sources, the elective category is the most frequent. If coded, newborn and delivery should each be around 10-13%. The number of newborn admissions should be close to the number of observations with AGEDAY = 0.
BILL	Health insurance billing number.	None.
BILL_S	Synthetic health insurance billing number.	None.

BWT	Birthweight in grams.	The number of records with nonmissing values should be close to the number with AGEDAY = 0.
		Expected mean = 3,300
		Minimum = 0
		Maximum = 65,535

CHG1-CHGnn	Charge detail.	Negative and zero values allowed.

DCCHPR1-DCCHPRnn	Clinical Classifications Software (CCS), formerly known as Clinical Classifications for Health Policy Research (CCHPR): Diagnosis classification.	The number of DCCHPRnn variables should correspond to the number of diagnoses provided by this data source.
		Minimum = 1
		Maximum = 260

DDATE	Discharge date.	Monthly frequencies should not fluctuate greatly. There may be some seasonal fluctuations, e.g., fewer admissions in the summer months.
DIED	Indicates in-hospital death.	Same number of missing values as DISP. Expected mean = 0.02-0.03 Minimum = 0 Maximum = 1 Number of records indicating died should match the number died under DISP.
DISP	Disposition of patient includes routine, short-term hospital, skilled nursing facility, intermediate facility, home health care, against medical advice, and death.	Hopefully, few missing values. Minimum = 1 Maximum = 20 Most records are routine discharges. Death rate should be 2-3%.

Table 1. HCUP Variables – Quality Review Guidelines		
Variable Name	**Description**	**Guidelines**
DOB	Date of birth.	Dates may be after the period of data if the birth century was erroneously reported. Since the distribution of age should be faintly trimodal with few values over 90, the distribution of DOB should also look trimodal. Because 10-13% of all admissions are births, approximately that many discharges will have a DOB in the discharge year.

DQTR	Discharge quarter.	Coded for all observations.
		Minimum = 0
		Maximum = 4
		Missing quarters are coded as 0. The number of records in each quarter should not fluctuate greatly.

DRG	DRG in use on discharge date.	Coded for all observations. Minimum = 1 Maximum varies by discharge date: Date Maximum 1/88-9/88 475 10/88-9/90 477 10/90-9/91 490 10/91-9/93 492 10/93-9/94 494 10/94-9/97 495 10/97-9/98 503 Percentage of records with DRG = 470 (ungroupable) should be small. If the percentage is greater than 5%, it may indicate a problem with the diagnoses and procedures. Confirm that the percentage of discharges with DRG = 469 (invalid principal diagnosis) is also small.

DRG10	DRG, Version 10.	Coded for all observations.
		Minimum = 1
		Maximum = 492
		Percent of records with DRG10 = 470 (ungroupable) should be small. If the percentage is greater than 5%, it may indicate a problem with the diagnoses and procedures. Confirm that the percent of discharges with DRG10 = 469 (invalid principal diagnosis) is also small. The percentage of records with DRG10 = 469 and 470 should be similar to those with DRG = 469 and 470.

Table 1. HCUP Variables – Quality Review Guidelines

Variable Name	Description	Guidelines
DRGVER	Grouper version in use on discharge date.	Coded for all observations. Minimum = 4 Maximum = 10 Frequency should be appropriate for the discharge date: Ver Effective Dates 4 10/1/87-9/30/88 5 10/1/88-9/30/89 6 10/1/89-9/30/90 7 10/1/90-9/30/91 9 10/1/91-9/30/92 10 10/1/92-9/30/93 11 10/1/93-9/30/94 12 10/1/94-9/30/95 13 10/1/95-9/30/96 14 10/1/96-9/30/97 15 10/1/97-9/30/98
DSHOSPID	Hospital number as received from the data source.	Coded for all observations.
DSNDX	Maximum number of diagnosis codes that could occur on a discharge record.	Coded for all observations.
DSNPR	Maximum number of procedure codes that could occur on a discharge record.	Coded for all observations.
DSNUM	Data source number.	Coded for all observations.
DSTYPE	Data source type indicates state data organization, hospital association, consortium, and other.	Coded for all observations.
DX1-DXnn	Diagnoses.	See next section: "Diagnosis and Procedure Code Variables."
DXSYS	Diagnosis coding system, usually ICD-9-CM.	Coded for all observations.

DXV1-DXVnn	Validity flag for diagnoses – indicates valid, invalid, or missing diagnosis.	The number of validity flags should correspond to the number of diagnoses provided by this data source.
		Expected Mean < 0.05
		Minimum = 0
		Maximum = 1

ED010-EDnnn	Edit-check variables – indicate inconsistencies among variables.	None.
HIC	Medicare beneficiary number.	None.
HIC_S	Synthetic Medicare beneficiary number.	None.

Table 1. HCUP Variables – Quality Review Guidelines

Variable Name	Description	Guidelines
HOSPID	HCUP hospital number.	In the SID, coded for all observations, 1988-1993. In the NIS, coded for all observations, regardless of year.
HOSPST	State postal code for hospital.	Coded for all observations.
HOSPSTCO	Modified FIPS state/county code for hospital.	In the SID, coded for all observations, 1988-1993. In the NIS, coded for all observations, regardless of year.
LOS	Length of stay, edited.	Missing at least as often as LOS_X. Expected mean = 4-10 Minimum = 0 Zero-day stays will account for 2-3% of admissions. Most stays will be less than one month. The distribution of the length of stay will vary greatly depending on two factors: the location of the hospital, and the type of services the hospital offers. East Coast hospitals tend to have longer stays (9-11 days) than West Coast hospitals (4-6 days). Hospitals with large rehabilitation, psychiatric, or long-term-care departments will have patients with extremely long stays. *(Note:* HCUP edits unjustifiably long stays over 365 days, and high or low charges per day). The distribution of length of stay should be right-skewed, some outliers should be expected, and the mean should be greater than the median.
LOS_X	Length of stay, unedited.	Missing less often than LOS. Minimum = 0 Maximum might be extreme. Unexplained long stays over 365 days and discharges with low or high charges per day have not been edited. The distribution should be right-skewed, some outliers should be expected, and the mean should be greater than the median.
MCDID	Medicaid recipient number.	None.
MCDID_S	Synthetic Medicaid recipient number.	None.

Table 1. HCUP Variables – Quality Review Guidelines		
Variable Name	**Description**	**Guidelines**

MDC	MDC in use on discharge date.	Coded for all observations.
		Minimum = 0
		Maximum = 25

MDC10	MDC, Version 10.	Coded for all observations.
		Minimum = 0
		Maximum = 25

MDID	Attending physician number as received from the data source.	None.
MDID_S	Synthetic attending physician number.	None.
MDNAME	Attending physician name.	None.
MDSPEC	Attending physician specialty.	None.
MRN	Medical record number as received from the data source.	None.
MRN_S	Synthetic medical record number.	None.
NDX	Number of diagnoses coded on the discharge record.	Coded for all observations. Minimum = 0 Maximum <= DSNDX There should be *few or no* records without diagnoses (NDX = 0). As the number of coded diagnoses increases, the corresponding proportion of discharges should decrease. (Percentages with NDX = 1 or 2 may be similar, though.) If the number of diagnoses supplied by the data source is low (e.g., <= 5), there may be a bulge at the maximum due to counting records that had at least that many diagnoses.

NEOMAT	Neonatal/maternal flag for neonatal diagnoses, maternal diagnoses/procedures, or both on a discharge record.	Coded for all observations.
		Minimum = 0
		Maximum = 3
		Percentage of maternal and neonatal should be similar, and approximately 10-13% each. Percentage of maternal records should be a bit higher than percentage of neonatal records. The number of combined neonatal/maternal records should match the number of records with ED100 = 1.

Table 1. HCUP Variables – Quality Review Guidelines

Variable Name	Description	Guidelines
NPR	Number of procedures coded on the discharge record.	Coded for all observations. Minimum = 0 Maximum <= DSNPR The percent of discharges without a procedure can vary from 20-50%. As the number of procedures coded increases, the corresponding proportion of discharges should decrease. If the number of procedures supplied by the data source is low (e.g., <= 5), there may be a bulge at the maximum due to counting records that had at least that many procedures.

PAY1	Expected primary payer includes Medicare, Medicaid, private insurance, self-pay, and no charge.	Hopefully, few missing values.
		Minimum = 1
		Maximum = 6
		Medicare will be 20-30%. Medicaid will be 5-15%.

PAY1_N	Expected primary payer (more detailed than PAY1) includes Medicare, Medicaid, Blue Cross, commercial, alternative delivery systems (e.g., HMO), self-pay, no charge, Title V, Workers' Compensation, CHAMPUS, and other government.	Missing as often as PAY1. Minimum = 1 Maximum = 12 Medicare will be 20-30%. Medicaid will be 5-15%. The percentage of Blue Cross will vary depending on the geographic area. East Coast hospitals have a much higher concentration of Blue Cross than West Coast hospitals. Consider the socioeconomic relationship between the hospital's clientele and the pay sources. Inner-city urban hospitals tend to treat a higher proportion of Medicaid and self-pay patients.
PAY1_X	Expected primary payer as received from the data source.	None.

PAY2	Expected secondary payer includes Medicare, Medicaid, private insurance, self-pay, and no charge.	May have many missing values.
		Minimum = 1
		Maximum = 6

PAY2_N	Expected secondary payer (more detailed than PAY2) includes Medicare, Medicaid, Blue Cross, commercial, alternative delivery systems, self-pay, no charge, Title V, Workers' Compensation, CHAMPUS, and other government.	Missing as often as PAY2.
		Minimum = 1
		Maximum = 12

Table 1. HCUP Variables – Quality Review Guidelines		
Variable Name	**Description**	**Guidelines**
PAY2_X	Expected secondary payer as received from the data source.	None.
PAY3_X	Expected tertiary payer as received from the data source.	None.

PCCHPR1-PCCHPRnn	Clinical Classifications Software (CCS), formerly known as Clinical Classifications for Health Policy Research (CCHPR): Procedure classification	The number of PCCHPRnn variables should correspond to the number of procedures provided by this data source.
		Minimum = 1
		Maximum = 231

PNUM	Person number as received from the data source.	None.
PNUM_S	Synthetic person number.	None.
PR1-PRnn	Procedures.	See next section: "Diagnosis and Procedure Code Variables."
PRDATE1-PRDATEnn	Date of procedure.	The number of date variables should correspond to the number of procedures provided by this data source. Missing at least as often as procedures.
PRDAY1	Days from admission of principal procedure.	The number of procedure-day variables should correspond to the number of procedures provided by this data source. Missing at least as often as procedures. Minimum = -4 Maximum = Maximum LOS+1 Highest frequency at 0. The frequency will fall off very rapidly after 0. The number of inconsistent PRDAY1 (negative 6-filled) should match the number of observations with ED701= 1 and ED801 = 1.
PROCESS	Processing number assigned for tracking records throughout data processing.	Coded on all observations.
PRSYS	Procedure coding system – ICD-9-CM, CPT-4, or HCPCS.	Coded on all observations.

PRV1-PRVnn	Validity flag for procedures – indicates valid, invalid, or missing procedure.	The number of validity flags should correspond to the number of procedures provided by this data source.
		Expected Mean < 0.05
		Minimum = 0
		Maximum = 1

PSTCO	Modified FIPS state/county code for patient.	None.

Table 1. HCUP Variables – Quality Review Guidelines		
Variable Name	**Description**	**Guidelines**
RACE	Race includes white, black, Hispanic, Asian, Pacific Islander, and Native American.	Can have numerous missing values. Minimum = 1 Maximum = 6 Check distributions against expectations, given the location of the hospitals. For example: • California may have high concentrations of Asian-Americans. • Areas with heavy urban concentrations (e.g., many states in the Northeast) would be expected to have high concentrations of blacks and other minorities. • Texas, California, and other Southwestern states, in addition to Florida, would be expected to have high concentrations of Hispanics.
RATE1-RATEnn	Charge detail expressed as the rate per unit (content varies by data source).	None.
RDRG	Refined DRG.	None.
RDRGWT	Refined DRG weight.	None.
READMIT	Readmission indicator.	None.
REVCD1-REVCDnn	Charge detail indicates revenue codes associated with detailed charges (content varies by data source).	None.
SEQ	Unique sequence number indicates the order in which the data file is sorted.	In the SID, coded on all observations, 1988-1993. In the NIS, coded for all observations, regardless of year.
SEQ_SID	Unique sequence number indicates the order in which the data file is sorted.	Coded on all observations. Included in database starting in 1994.
SEX	Sex includes male or female.	Few missing values. Expected mean = 1.6 Minimum = 1 Maximum = 2 Expect 60% female. Even excluding deliveries, females tend to be more frequent users of medical services.
SURGID	Primary surgeon number as received from the data source.	None.

Table 1. HCUP Variables – Quality Review Guidelines

Variable Name	Description	Guidelines
SURGID_S	Synthetic primary surgeon number.	None.

TMDX1-TMDXn	Time of onset for each diagnosis indicates whether the diagnosis was present at admission.	Missing at least as often as diagnosis.
		Minimum = 0
		Maximum = 1

TOTCHG	Total charges, edited.	Missing at least as often as TOTCHG_X. No zero values allowed. Dollars rounded. Expected Mean = 3,000-10,000 Minimum = 1 Maximum = 9,999,999,999 The distribution of total charges is very sensitive to the length of stay. Hospitals that on average have long lengths of stay will have higher total charges. *(Note:* HCUP edits high or low charges per day). The distribution of total charges should be right-skewed, some outliers should be expected, and the mean should be greater than the median.
TOTCHG_X	Total charges, unedited.	Hopefully, few missing values. No zero values allowed. Retains cents if supplied. Minimum = -9,999,999,999.99 Maximum = +9,999,999,999.99 Discharges with unjustifiably high or low charges per day have not been edited. The distribution should be right-skewed, some outliers should be expected, and the mean should be greater than the median.
UNIT1-UNITnn	Charge detail expressed as number of units of specific services (content varies by data source).	All values should be rounded to the nearest whole dollar. Minimum = 1
YEAR	Discharge year (calendar).	Coded for all observations.
ZIP	Zip code of patient.	None.
ZIP_S	Synthetic zip code of patient.	None.

ZIPINC4	Median household income of patient's zip code (4 categories).	Missing as often as ZIP.
		Minimum = 1
		Maximum = 4

Table 1. HCUP Variables – Quality Review Guidelines		
Variable Name	**Description**	**Guidelines**

ZIPINC8	Median household income of patient's zip code (8 categories).	Missing as often as ZIP and at least as often as ZIPINC4.
		Minimum = 1
		Maximum = 8
		If there are fewer than 3 zip codes within any ZIPINC8 income category for a state, ZIPINC8 is set to missing for the entire state. Only ZIPINC4 will be available for these zip codes.

DIAGNOSIS AND PROCEDURE CODE VARIABLES

The coding of the diagnosis/procedure-specific variables is interdependent. These variables are:

- Diagnoses (DX1-DX30) and procedures (PR1-PR30)
- Validity flags (DXV1-DXV30 and PRV1-PRV30)
- CCS codes (DCCHPR1-DCCHPR30 and PCCHPR1-PCCHPR30)

Table 2 demonstrates the relationship between these variables.

Table 2. Relationship Between Diagnosis and Procedure Codes and Their Associated Variables

Diagnosis (DXn)/ Procedure Code (PRn)	Validity Flags DXVn/PRVn	CCS Codes DCCHPRn/PCCHPRn
Missing	. or -9	. or -999
Valid Code	0 if consistent with age and sex; .C or -6 if inconsistent	1-260/ 1-231
Invalid Code	1	.A or -888

HCUP EDIT CHECKS

Table 3 lists all of the edit checks performed on the HCUP inpatient discharge data, along with their associated variable names. Variables are named EDnnn, where nnn is a unique number. HCUP uses many diagnosis and procedure screens to define specific conditions employed in the editing procedures. These screens are defined following the edit-check table. The condition column specifies the source code used to identify the problem. Variables with the prefix "I." contain information as provided by the data source (e.g., I.LOS is length of stay as provided by the data source).

Table 3. HCUP Edit Checks

Edit Check	Description	Condition	Action
ED010	**REPORTED LOS IS NOT EQUAL TO CALCULATED LOS** The length of stay calculated from admission date and discharge date does not equal the reported length of stay.	I.LOS ne LOS_X	For tabulation purposes only.
ED011	**ADMIT DATE IS AFTER DISCHARGE DATE** The length of stay is negative.	LOS < 0	Set ADATE and LOS to inconsistent (.C).
ED020	**REPORTED AGE IN YEARS DOES NOT EQUAL CALCULATED AGE** The age in years calculated from birthdate and admission date does not equal the reported age.	I.AGE ne AGE	For tabulation purposes only.
ED021	**AGE IN YEARS INCONSISTENT WITH INFANT AGE** Infant age is nonmissing, but the age in years is greater than zero.	(AGEDAY >= 0) and (AGE > 0)	Set AGEDAY and AGE to inconsistent (.C).
ED100	**MATERNAL AND NEONATAL RECORD** Codes in the diagnosis vector or the procedure vector satisfy both the maternal and neonatal screens.	DX1-DX30 or PR1-PR30 are MATERNAL and NEONATE	For tabulation purposes only.
ED101	**PRINCIPAL DIAGNOSIS INCONSISTENT WITH SEX** The sex coded for the patient does not agree with the sex of the principal diagnosis.	SEX ne Sex of DX1	Set DXV1 and SEX to inconsistent (.C).
ED102-ED1nn	**SECONDARY DIAGNOSIS INCONSISTENT WITH SEX** The sex coded for the patient does not agree with the sex of a secondary diagnosis.	SEX ne Sex of DXn	Set DXVn and SEX to inconsistent (.C).
ED201	**PRINCIPAL PROCEDURE INCONSISTENT WITH SEX** The sex coded for the patient does not agree with the sex of the principal procedure.	SEX ne Sex of PR1	Set PRV1 and SEX to inconsistent (.C).

Table 3. HCUP Edit Checks

Edit Check	Description	Condition	Action
ED202-ED2nn	**SECONDARY PROCEDURE INCONSISTENT WITH SEX** The sex coded for the patient does not agree with the sex of a secondary procedure.	SEX ne Sex of PRn	Set PRVn and SEX to inconsistent (.C).
ED301	**NEONATAL PRINCIPAL DIAGNOSIS INCONSISTENT WITH AGE** The principal diagnosis satisfies the NEONATE screen, and the age in years is greater than zero. Retain age on a combined neonatal/maternal record.	(DX1 is NEONATE) and (AGE > 0)	Set DXV1 to inconsistent (.C). If NEOMAT ne 3, set AGE and AGEDAY to inconsistent (.C).
ED302-ED3nn	**NEONATAL SECONDARY DIAGNOSIS INCONSISTENT WITH AGE** A secondary diagnosis satisfies the NEONATE screen, and the age in years is greater than zero. Retain age on a combined neonatal/maternal record.	(DXn is NEONATE) and (AGE > 0)	Set DXVn to inconsistent (.C). If NEOMAT ne 3, set AGE and AGEDAY to inconsistent (.C).
ED401	**MATERNAL PRINCIPAL DIAGNOSIS INCONSISTENT WITH AGE** The principal diagnosis satisfies the MATERNAL screen, and the nonmissing age in years is less than 10 or greater than 55. Retain age on a combined maternal/neonatal record.	(DX1 is MATERNAL) and NOT (10 <= AGE <= 55)	Set DXV1 to inconsistent (.C). If NEOMAT ne 3, set AGE and AGEDAY to inconsistent (.C).
ED402-ED4nn	**MATERNAL SECONDARY DIAGNOSIS INCONSISTENT WITH AGE** A secondary diagnosis satisfies the MATERNAL screen, and the nonmissing age in years is less than 10 or greater than 55. Retain age on a combined maternal/neonatal record.	(DXn is MATERNAL) and NOT (10 <= AGE <= 55)	Set DXVn to inconsistent (.C). If NEOMAT ne 3, set AGE and AGEDAY to inconsistent (.C).

Table 3. HCUP Edit Checks

Edit Check	Description	Condition	Action
ED501	**MATERNAL PRINCIPAL PROCEDURE INCONSISTENT WITH AGE** The principal procedure satisfies the MATERNAL screen, and the nonmissing age in years is less than 10 or greater than 55. Retain age on a combined maternal/neonatal record.	(PR1 is MATERNAL) and NOT (10 <= AGE <= 55)	Set PRV1 to inconsistent (.C). If NEOMAT ne 3, set AGE and AGEDAY to inconsistent (.C).
ED502-ED5nn	**MATERNAL SECONDARY PROCEDURE INCONSISTENT WITH AGE** A secondary procedure satisfies the MATERNAL screen, and the nonmissing age in years is less than 10 or greater than 55. Retain age on a combined maternal/neonatal record.	(PRn is MATERNAL) and NOT (10 <= AGE <=55)	Set PRVn to inconsistent (.C). If NEOMAT ne 3, set AGE and AGEDAY to inconsistent (.C).
ED600	**LONG LOS, JUSTIFIED** The length of stay is over 365 days, and is justified by a long-term-care diagnosis, a perinatal diagnosis, discharge to another facility, or the patient's death.	(LOS > 365) and (PERINATE or LTC or (2 <= DISP <= 5) or (DIED = 1))	For tabulation purposes only.
ED601	**LONG LOS, UNJUSTIFIED** The length of stay is over 365 days, and is not justified by a long-term-care diagnosis, a perinatal diagnosis, discharge to another facility, or the patient's death.	(LOS > 365) and NOT (PERINATE or LTC or (2 <= DISP <= 5) or (DIED = 1))	Set LOS to inconsistent (.C).
ED701	**DAY OF PRINCIPAL PROCEDURE WITHOUT PRINCIPAL PROCEDURE** There is a nonmissing day of principal procedure without a corresponding principal procedure.	(PRDATE1 or PRDAY1 ne .) and (PR1 = ' ')	Set PRDAY1 and PRDATE1 to inconsistent (.C).
ED702-ED7nn	**DAY OF SECONDARY PROCEDURE WITHOUT CORRESPONDING PROCEDURE** There is a nonmissing day of secondary procedure without a corresponding procedure code.	(PRDATEn or PRDAYn ne .) and (PRn = ' ')	Set PRDAYn and PRDATEn to missing (.) and move up all subsequent procedure date pairs.

Table 3. HCUP Edit Checks

Edit Check	Description	Condition	Action
ED801	**DAY OF PRINCIPAL PROCEDURE NOT DURING STAY** The nonmissing day of the principal procedure is less than (-4) or greater than the length of stay plus one.	NOT (-4 <= PRDAY1 <= LOS+1)	Set PRDAY1 and PRDATE1 to inconsistent (.C).
ED802-ED8nn	**DAY OF SECONDARY PROCEDURE NOT DURING STAY** The nonmissing day of secondary procedure is less than (-4) or greater than the length of stay plus one.	NOT (-4 <= PRDAYn <= LOS+1)	Set PRDAYn and PRDATEn to inconsistent (.C).
ED910	**CHARGES PER DAY ARE EXCESSIVELY LOW, JUSTIFIED** Total charges and length of stay are both nonmissing; charges per day are less than $100, and are justified by discharge to another facility or by the patient's death.	(TOTCHG ÷ LOS < 100) and ((2 <= DISP <= 5) or (DIED = 1))	For tabulation purposes only.
ED911	**CHARGES PER DAY ARE EXCESSIVELY LOW, UNJUSTIFIED** Total charges and length of stay are both nonmissing; charges per day are less than $100, and are not justified by discharge to another facility or by the patient's death.	(TOTCHG ÷ LOS < 100) and NOT ((2 <= DISP <= 5) or (DIED = 1))	Set TOTCHG and LOS to inconsistent (.C).
ED920	**CHARGES PER DAY ARE EXCESSIVELY HIGH, JUSTIFIED** Total charges and length of stay are both nonmissing; charges per day are more than $20,000, and are justified by discharge to another facility or by the patient's death.	(TOTCHG ÷ LOS > 20000) and ((2 <= DISP <= 5) or (DIED = 1))	For tabulation purposes only.
ED921	**CHARGES PER DAY ARE EXCESSIVELY HIGH, UNJUSTIFIED** Total charges and length of stay are both nonmissing; charges per day are more than $20,000, and are not justified by discharge to another facility or by the patient's death.	(TOTCHG ÷ LOS > 20000) and NOT ((2 <= DISP <= 5) or (DIED = 1))	Set TOTCHG and LOS to inconsistent (.C).
ED951	**UNACCEPTABLE UNIFORM PAY SOURCE COMBINATION** The uniform primary pay source and secondary pay source are the same, and the sources are Medicare or Medicaid.	(PAY1 = PAY2) and (1 <= PAY2 <= 2)	Set PAY2 and PAY2_N to inconsistent (.C).
ED952	**UNACCEPTABLE NON-UNIFORM PAY SOURCE COMBINATION** The non-uniform primary pay source and secondary pay source are the same, and the sources are CHAMPUS, Worker's Compensation, or Title V.	(PAY1_N = PAY2_N) and (8 <= PAY2_N <= 10)	Set PAY2 and PAY2_N to inconsistent (.C).

DIAGNOSIS AND PROCEDURE SCREENS

The diagnosis and procedure screens used in HCUP inpatient discharge data processing are specified below. Codes added because of changes in ICD-9-CM coding are underlined.

Maternal: Screen used for 1988 to 1993 data:
Diagnoses 630 to 67694; V220 to V242; and V270 to V279
Procedures 720 to 7599

Screen used for 1994 to 1996 data:
Diagnoses 630 to 677; V220 to V242; and V270 to V279
Procedures 720 to 7599;

Screen used beginning in 1997:
Diagnoses 630 to 677; V220 to V242; and V270 to V279
Procedures 720 to 7599; 7965

Neonate: Screen used for 1988 to 1993 data:
Diagnoses 7600 to 7799; and V3000 to V392

Screen used for 1994-1995:
Diagnoses 75983, 7600 to 7799, V3000 to V392

Note: Code 75983 was erroneously included in the neonate screen. Because this is a rare condition, only a negligible number of records should be affected.

Screen used beginning in 1996:
Diagnoses 7600 to 7799, V3000 to V392

Perinate: Screen used beginning in 1988:
Diagnoses 7400 to 7799

Long-term-care indication:

Screen used for 1988 to 1992 data:
Diagnoses 2900 to 30503; 30520 to 3124; 3219 to 319; 3440; 430 to 438; and 797 to 7999

Screen used for 1993 data:
Diagnoses 2900 to 30503; 30520 to 3124; 3219 to 319; 3440; 34481; 430 to 438; 44024, 4416, 78003, and 797 to 7999

Note: Codes 78001, 78002, and 78009 were erroneously excluded from the long-term care screen. This would cause some discharges with long length of stays (over 365 days) to have ED601 "Long Length of Stay, Unjustified" set instead of ED600 "Long Length of Stay, Justified."

Screen used for 1994 data:

Diagnoses 2900 to 30503, 30520 to 3124, 3129 to 319, 34400 to 34409, 34481, 430 to 438, 44024, 4416, 78003, 797 to 7999

Note: Codes 78001, 78002, 78009, 31281, 31282, and 31289 were erroneously excluded from the long-term care screen. This would cause some discharges with long length of stays (over 365 days) to have ED601 "Long Length of Stay, Unjustified" set instead of ED600 "Long Length of Stay, Justified."

Screen used for 1995 data:

Diagnoses 2900 to 30503, 30520 to 3124, 3129 to 319, 34400 to 34409, 34481, 430 to 4352, 4358 to 438, 44024, 4416, 78003, 797 to 7999

Note: Codes 78001, 78002, 78009, 31281, 31282, 31289, and 4353 were erroneously excluded from the long-term care screen. This would cause some discharges with long length of stays (over 365 days) to have ED601 "Long Length of Stay, Unjustified" set instead of ED600 "Long Length of Stay, Justified."

Screen used for 1996 data:

Diagnoses 2900 to 319, 34400 to 34409, 34481, 430 to 438, 44024, 4416, 78001 to 78009, 797 to 7999

Screen used beginning in 1997:

Diagnoses 2900 to 319, 34400 to 34409, 34481, 430 to 4389, 44024, 4416, 78001 to 78009, 797 to 7999

Male diagnoses:

Screen used for 1988 to 1992 data:

Diagnoses 01640 to 01656, 05413, 0720, 09812 to 09814, 09832 to 09834, 13103, 1750 to 1759, 185 to 1879, 2144, 2220 to 2229, 2334 to 2336, 2364 to 2366, 2570 to 2579, 30274 to 30275, 4564, 600 to 6089, 7525 to 7526, 7587, 78832, 7922, 8780 to 8783, 9393, V1045 to V1049, V502

Screen used for 1993 to 1995 data:

Diagnoses 01640 to 01656, 05413, 0720, 09812 to 09814, 09832 to 09834, 13103, 1750 to 1759, 185 to 1879, 2144, 2220 to 2229, 2334 to 2336, 2364 to 2366, 2570 to 2579, 30274 to 30275, 4564, 600 to 6089, 7525 to 7526, 7587, 78832, 79093, 7922, 8780 to 8783, 9393, V1045 to V1049, V502

Screen used in 1996 data:

Diagnoses 01640 to 01656, 05413, 0720, 09812 to 09814, 09832 to 09834, 13103, 1750 to 1759, 185 to 1879, 2144, 2220 to 2229, 2334 to 2336, 2364 to 2366, 2570 to 2579, 30274 to 30275, 4564, 600 to 6089, 75251 to 75269, 7587, 78832, 79093, 7922, 8780 to 8783, 9393, V1045 to V1049, V502

Screen used beginning in 1997:

Diagnoses 01640 to 01656, 05413, 0720, 09812 to 09814, 09832 to 09834, 13103, 1750 to 1759, 185 to 1879, 2144, 2220 to 2229, 2334 to 2336, 2364 to 2366, 2570 to 2579, 30274 to 30275, 4564, 600 to 6089, 75251 to 75269, 7587, 78832, 79093, 7922, 8780 to 8783, 9393, V1045 to V1049, V1642-V1643, V502

Male procedures:

Screen used beginning in 1988:

Procedures 600 to 6499, 8791 to 8799, 9824, 9994 to 9996.

Female diagnoses:

Screen used from 1988 to 1995:

Diagnoses 01660 to 01676, 05411 to 05412, 09815 to 09817, 09835 to 09837, 1121, 13101, 1740 to 1749, 179 to1849, 1986, 2180 to 2219, 2331 to 2333, 2360 to 2363, 2560 to 2569, 30273, 30276, 30651 to 30652, 4566, 6115 to 6116, 6140 to 66942, 66944 to 67694, 71630 to 71639, 7520 to 75249, 7923, 7950, 8674 to 8675, 8784 to 8787, 90255 to 90256, 90281 to 90282, 9391 to 9392, 9474, 99632, V074, V1040 to V1044, V131, V220 to V235, V238 to V2501, V251, V253, V2541 to V2543, V255, V261, V270 to V289, V447, V524, V557, V723 to V724, V762

Note: Starting in 1994, Codes 66943, 677, V237, V4551, V4552, and V5042 were erroneously excluded from the female screen. This would cause ED1nn "Diagnosis Inconsistent with Sex" to not be set when a male discharge had one of these female diagnoses.

Screen used in 1996:

Diagnoses 01660 to 01676, 05411 to 05412, 09815 to 09817, 09835 to 09837, 1121, 13101, 1740 to 1749, 179 to1849, 1986, 2180 to 2219, 2331 to 2333, 2360 to 2363, 2560 to 2569, 30273, 30276, 30651 to 30652, 4566, 6115 to 6116, 6140 to 677, 71630 to 71639, 7520 to 75249, 7923, 7950, 8674 to 8675, 8784 to 8787, 90255 to 90256, 90281 to 90282, 9391 to 9392, 9474, 99632, V074, V1040 to V1044, V131, V220 to V2501, V251, V253, V2541 to V2543, V255, V261, V270 to V289, V447, V4551-V4552, V5042, V524, V557, V723 to V724, V762

Note: Code E9672 was erroneously included in the female screen when processing 1996 data for all states and 1997 for a few states. This would cause male discharges with the diagnosis E9672 “Child and adult battering and other maltreatment -- by mother or step mother” to have edit check ED1nn set to 1 and the diagnosis validity flag DXVn and SEX set to inconsistent (.C). Because this is a rarely used code, only a negligible number of records should be affected.

Screen used beginning in 1997:

Diagnoses 01660 to 01676, 05411 to 05412, 09815 to 09817, 09835 to 09837, 1121, 13101, 1740 to 1749, 179 to1849, 1986, 2180 to 2219, 2331 to 2333, 2360 to 2363, 2560 to 2569, 30273, 30276, 30651 to 30652, 4566, 6115 to 6116, 6140 to 677, 71630 to 71639, 7520 to 75249, 7923, 7950, 7965, 8674 to 8675, 8784 to 8787, 90255 to 90256, 90281 to 90282, 9391 to 9392, 9474, 99632, V074, V1040 to V1044, V131, V1641, V220 to V2501, V251, V253, V2541 to V2543, V255, V261, V270 to V289, V447, V4551-V4552, V5042, V524, V557, V723 to V724, V762

Note: Code E9672 was erroneously included in the female screen when processing 1996 data for all states and 1997 for a few states. This would cause male discharges with the diagnosis E9672 “Child and adult battering and other maltreatment -- by mother or step mother” to have edit check ED1nn set to 1 and the diagnosis validity flag DXVn and SEX set to inconsistent (.C). Because this is a rarely used code, only a negligible number of records should be affected.

Female procedures:

Screen used for 1988 to 1995 data:

Procedures 650 to 7599, 8781 to 8789, 8846, 8878, 8926, 9141 to 9149, 9217, 9614 to 9618, 9644, 9724 to 9726, 9771 to 9775, 9816 to 9817, 9823, 9998

Screen used beginning in 1996:

Procedures 6501 to 7599, 8781 to 8789, 8846, 8878, 8926, 9141 to 9149, 9217, 9614 to 9618, 9644, 9724, 9726, 9771 to 9775, 9816 to 9817, 9823, 9998

TECHNICAL SUPPLEMENT 3:
MAPPING SOURCE-SPECIFIC HOSPITAL IDENTIFIERS
TO AHA HOSPITAL IDENTIFIERS

INTRODUCTION

The American Hospital Association (AHA) definition of "community hospital" was used to select hospitals for the HCUP databases. Therefore, for each participating data source and for each year, it was necessary to reconcile the data source's identification of the hospital with the identification of the hospital in the associated AHA Annual Survey. The list of all such linkages is called a *crosswalk*.

Once these linkages were established, data from the AHA Annual Survey were used to:

- identify facilities that are defined by the AHA as "community hospitals," which are therefore eligible for inclusion in the HCUP inpatient databases;

- classify each community hospital into a stratum for sampling for the Nationwide Inpatient Sample (NIS);

- add various types of information about the hospital (such as its county FIPS code) to the inpatient records; and

- identify community hospitals listed in the AHA Annual Survey for which no inpatient data were supplied by the data source.

This Technical Supplement addresses the procedures used to identify the appropriate linkages between AHA hospital identifiers and hospitals represented in the inpatient data supplied by each data source.

Note: In this document, *data source* always refers to the source of the inpatient data.

RECONCILING HOSPITAL IDENTIFIERS

The goal is to identify an appropriate AHA hospital identifier for each source hospital.

To begin, relevant data elements are extracted from inpatient data and from the AHA data for each year. The two elements extracted from the inpatient data are:

- the source-specific hospital identifier, and

- a count of the hospital's inpatient records for each quarter and for the year.

Electronic Linkage of Source and AHA Hospital Identifiers

First, the hospital identifiers used by the data source (hereafter referred to as DSHOSPID) are linked electronically to the relevant data elements extracted from the AHA Annual Survey data. AHA identifiers include all hospitals in the state, not just the community hospitals that are eligible for inclusion in HCUP.

A SAS merge step is used to link the DSHOSPIDs to AHA identifiers. The specific variables used in the merge depend on the information provided by the data source. In order of preference, these variables are:

- hospital name, city, and zip code;
- hospital name; or
- any other unique variable that is available – e.g., Medicare provider number.

The AHA and the data source often use different methods to represent components of a hospital's name (e.g., "Community General Hospital" may be represented as "Community General Hosp" by the AHA and as "Community Gen Hosp" by the data source). Hence, before the SAS merge, the AHA and the source's hospital names are transformed into a uniform link variable, which imposes similar methods of abbreviations, lowercase and uppercase letters, and different characters. This effectively reduces the number of non-merges that occur simply because of different methods of representing the hospital name's components.

Three types of linkages result from this step, as shown in Table 4.

Table 4. Linkages Between AHA and Data Source Identifiers

Row	AHA Identifier	DSHOSPID	Link?
1	present	present	yes
2	present	absent	no
3	absent	present	no

Approximately 80 percent of DSHOSPIDs link to AHA identifiers in this step. (These successful links are represented by Row 1 in Table 4). The other 20 percent of DSHOSPIDs (Rows 2 and 3) must be linked manually, using the process described below.

Prior experience has shown that a large majority of the hospitals failing to link in this step will usually be of the following types:

- closures,
- openings (new hospitals),
- mergers,
- demergers,
- dates of changes in the hospital structure that differ between the data source and the AHA Annual Survey, and

- levels of aggregation that differ between the data source and the AHA Annual Survey (e.g., the data source treats two separate facilities as two hospitals, while the AHA Annual Survey treats the two facilities as a single hospital, or vice versa).

Resolution of Unmatched Hospitals

The goal in this step is to identify an appropriate AHA hospital identifier for each source hospital that was not matched electronically above.

Several external sources of information are used to reconcile the unmatched source hospitals (Row 3) and the unmatched AHA hospitals (Row 2). These are:

- *Source Documentation:* This information, received from the data source, usually contains a list of hospitals, their cities (specific addresses are not always included), and the source's hospital identifier. This documentation is the primary source for finding missing information (e.g., specific names and addresses) for an unmatched hospital.

- *AHA Summaries:* The AHA Summary of Registered Hospitals and the AHA Summary of Nonregistered Hospitals, which are usually delivered with the annual *AHA Guide*, document additions and deletions reflected in the hospitals' responses to the AHA Annual Survey.

- *AHA Guide:* The *AHA Guide*, an annual hard-copy volume published by the AHA, provides a wealth of information about registered hospitals. The *AHA Guide* includes an entire section on multihospital health-care systems that identifies the hospitals included in specific multihospital systems. The *AHA Guide* also provides information about individual hospitals (organized by state, and within each state, by city), which includes:

    - Information also available from the AHA Annual Survey data; for example:
        - type of service (general medical/surgical, rehabilitation);
        - average lengths of stay (long- or short-term);
        - type of ownership; and
        - numbers of beds, admissions, births, etc.

    - Information about hospitals embedded within the organizational structure of another hospital; for example:

        Binghamton – Broome County, NY
            United Health Services (includes Binghamton General Hospital, Mitchell Ave. ...; Wilson Memorial Hospital, ... Harrison St. ...)

    - Information about changes in a hospital's name; for example:

        Dobbs Ferry – Westchester County, NY
            Community Hospital at Dobbs Ferry (formerly Dobbs Ferry Hospital)

    - References to a new hospital name or new location; for example:

        Delhi – Delaware County, NY
            A. Lindsay and Olive B. O'Connor Hospital. See Mary Imogene Bassett Hospital, Cooperstown.

- *Record counts generated from the supplied inpatient data:*
  The number of discharges reported in the inpatient data is compared to the number of discharges reported to the AHA. While this information is rarely definitive in linking source identifiers to AHA identifiers, it is sometimes useful in identifying a link to an AHA hospital, and provides a means of validating linkages obtained by other means. This information is especially useful in distinguishing a single hospital from two combined hospitals.

When it is not clear what needs to be done to a hospital or group of hospitals, an AHCPR analyst is consulted.

The reconciliation process is complete when the following facts are confirmed:

- all source hospital identifiers have been assigned an HCUP hospital identifier (HOSPID), and

- all hospitals (HOSPID) have only one FIPS county code assigned.

  Hospitals composed of multiple facilities in different locations are assigned the FIPS county code of the major facility, as defined by the AHA.

RULES FOR RESOLVING PROBLEM HOSPITALS

Following are the rules used for resolving problem hospitals. In these examples, the HCUP hospital identifier (HOSPID) starts with "SS" to indicate the state FIPS code:

The HCUP hospital identifier (HOSPID) reflects the AHA view of a hospital and is a randomly assigned number based on the AHA hospital identifier (IDNUMBER). If the data source reports the data from facilities that are combined in the AHA hospital definition, the IDNUMBER and the HOSPID will be the same for all the facilities. In the following example, three different source identifiers are considered to be part of one facility as defined by the AHA:

Year	*Data Source*	*AHA IDNUMBER*	*HOSPID*
1990	165 (Acute Care Unit)	910140	SS089
1990	165P (Psychiatric Unit)	910140	SS089
1990	166S (Swing Bed Unit)	910140	SS089

Openings

The AHA IDNUMBER and HCUP HOSPID are assigned to a newly opened hospital only when the hospital has first been recognized by the AHA for a particular survey year, even if the data source supplies data for an earlier time period. For example, the data source supplied data for a hospital starting in 1989, but the AHA first recognized the hospital in 1991:

Year	*Data Source*	*AHA IDNUMBER*	*HOSPID*
1989	86-0601625		
1990	86-0601625		
1991	86-0601625	860001	SS014
1992	86-0601625	860001	SS014

Closures

When a hospital closes (in the AHA's view), the AHA IDNUMBER and HCUP HOSPID are carried forward if there are inpatient data available from the data source. In this example, the AHA considered the hospital closed in 1990, but the data source still supplied data:

Year	*Data Source*	*AHA IDNUMBER*	*HOSPID*
1988	047	450520	SS171
1989	047	450520	SS171
1990	047	450520 (closed)	SS171

Mergers

When two or more hospitals merge (in the AHA's view), the IDNUMBER (along with the HOSPID) of the merged entity is assigned to all its component hospitals even if they continue reporting separately to the state. In this example, two hospitals have different source identifiers, but starting in 1990 are considered one facility by the AHA because of a merger:

Year	*Data Source*	*AHA IDNUMBER*	*HOSPID*
1989	036	450400	SS091
1990	036	450002 (merger)	SS013
1991	036	450002 (merger)	SS013
1992	036	450002 (merger	SS013
1989	126	451750	SS169
1990	126	450002 (merger)	SS013
1991	126	450002 (merger)	SS013
1992	126	450002 (merger)	SS013

Demergers

When hospitals demerge (in the AHA's view), the component hospitals are assigned a new AHA IDNUMBER or the one they previously had. The HCUP HOSPID follows the AHA IDNUMBER, so that the HOSPID changes if the IDNUMBER changes and the HOSPID is reused if the IDNUMBER is reused. In this example, a hospital demerges in 1989 into two facilities:

Year	*Data Source*	*AHA IDNUMBER*	*HOSPID*
1988	562	220515 (merger)	SS051
1989	562	220547 (demerger)	SS026
1990	562	220547 (demerger)	SS026
1988	561	220515 (merger)	SS051
1989	561	221240 (demerger)	SS037
1990	561	221240 (demerger)	SS037

Changes in Hospital Characteristics

(Note: The following decision is made only after AHCPR is consulted.) If during HCUP processing of the inpatient data, summary statistics on the distribution of length of stay look questionable for a community hospital (e.g., the mean length of stay is considerably greater than 30 days), the AHA community flag is investigated. If the AHA community flag was imputed from previous years because a hospital did not report to the AHA – and the data source can confirm that the facility is no longer a community hospital – the AHA identifier is still assigned to the facility, the community flag is imputed, and the hospital is excluded from the HCUP inpatient databases. In this example, the facility was considered a noncommunity hospital starting in 1990:

Year	*Data Source*	*AHA IDNUMBER*	*HOSPID*	*Community Flag*
1988	86-0201864	860575	SS090	1
1989	86-0201864	860575	SS090	1
1990	86-0201864	860575	SS101	0 (changed)*
1991	86-0201864	860575	SS101	0 (changed)*

*Hospitals with the community flag indicator of "0" are not processed.

TECHNICAL SUPPLEMENT 4: SOURCES OF HCUP DATA

Arizona
Arizona Department of Health Services

California
California Office of Statewide Health Planning and Development

Colorado
Colorado Health & Hospital Association

Connecticut
CHIME: Connecticut Health Information Management and Exchange

Florida
Florida Agency for Health Care Administration

Georgia
GHA: An Association of Hospitals and Health Systems

Hawaii
Hawaii Health Information Corporation

Illinois
Illinois Health Care Cost Containment Council

Iowa
Association of Iowa Hospitals and Health Systems

Kansas
Kansas Hospital Association

Maryland
Maryland Health Services Cost Review Commission

Massachusetts
Massachusetts Division of Health Care Finance and Policy

Missouri
Hospital Industry Data Institute

New Jersey
New Jersey Department of Health and Senior Services

New York
New York State Department of Health

Oregon
Oregon Association of Hospitals and Health Systems
Office for Oregon Health Plan Policy and Research

Pennsylvania
Pennsylvania Health Care Cost Containment Council

South Carolina
South Carolina State Budget and Control Board

Tennessee
THA: An Association of Hospitals and Health Systems

Utah
Utah Department of Health

Washington
Washington State Department of Health

Wisconsin
Wisconsin Department of Health and Family Services

Hospital Data
American Hospital Association

Zip Code Data
CACI Marketing Systems

TECHNICAL SUPPLEMENT 5:
DESIGN OF THE HCUP NATIONWIDE INPATIENT SAMPLE, RELEASE 1

INTRODUCTION

The Nationwide Inpatient Sample (NIS) of the Healthcare Cost and Utilization Project (HCUP) was established to provide analyses of hospital utilization across the United States. Release 1 covers calendar years 1988 through 1992. The target universe includes all acute-care discharges from all community hospitals in the United States; the NIS comprises all discharges from a sample of hospitals in this target universe.

For each calendar year (1988 through 1992), this first release of the NIS contains 5.2 to 6.2 million discharges from a sample of 758 to 875 hospitals in 11 states (8 states for 1988). Future releases will add data to the NIS file from more states and more years. Thus, the NIS supports both cross-sectional and longitudinal analyses.

Potential research issues focus on both discharge- and hospital-level outcomes. Discharge outcomes of interest include trends in inpatient treatments with respect to:

- frequency,
- costs,
- lengths of stay,
- effectiveness,
- appropriateness, and
- access to hospital care.

Hospital outcomes of interest include:

- mortality rates,
- complication rates,
- patterns of care,
- diffusion of technology, and
- trends toward specialization.

These and other outcomes are of interest for the nation as a whole and for policy-relevant inpatient subgroups defined by geographic regions, patient demographics, hospital characteristics, physician characteristics, and pay sources.

This report provides a detailed description of the NIS sample design, as well as a summary of the resultant hospital sample. Sample weights were developed to obtain national estimates of hospital and inpatient parameters. These weights and other special-use weights are described in detail.

THE NIS HOSPITAL UNIVERSE

For each calendar year, the hospital universe is defined by all hospitals that were open during any part of that calendar year and were designated as community hospitals in the American Hospital Association (AHA) Annual Survey of Hospitals. For purposes of the NIS, the definition of a community hospital is that used by the AHA: "all nonfederal short-term general and other specialty hospitals, excluding hospital units of institutions." Consequently, Veterans Hospitals and other federal hospitals are excluded.

Table 5 shows the number of universe hospitals for each calendar year based on HCUP's calendar-year conforming version of the AHA survey-year files. Survey responses were put on a calendar-year basis for 1988-1991 by merging data from adjacent survey years. However, 1992 AHA survey data remain in the original reporting-year form because HCUP received the 1993 AHA files too late to convert fiscal-year responses to calendar-year files for 1992.

Table 5. Hospital Universe

Calendar Year	Number of Hospitals
1988	5,607
1989	5,548
1990	5,468
1991	5,412
1992	5,334

Hospital Merges, Splits, and Closures

All hospital entities that were designated community hospitals in the AHA hospital file were included in the hospital universe. Therefore, if two or more community hospitals merged to create a new community hospital, the original hospitals and the newly-formed hospital were all considered separate hospital entities in the universe for the year of the merge. Likewise, if a community hospital split, the original hospital and all newly created community hospitals were separate entities in the universe for the year of the split. Finally, community hospitals that closed during a year were included as long as they were in operation during some part of the calendar year.

Stratification Variables

To help ensure representativeness, sampling strata were defined based on five hospital characteristics contained in the AHA hospital files. The stratification variables were as follows:

1) *Geographic Region – Northeast, North Central, West, and South*. This is an important stratifier because practice patterns have been shown to vary substantially by region. For example, lengths of stay tend to be longer in East Coast hospitals than in West Coast hospitals.

2) *Ownership – public, private not-for-profit, and private investor-owned.* These types of hospitals tend to have different missions and different responses to government regulations and policies.

3) *Location – urban or rural.* Government payment policies often differ according to this designation. Also, rural hospitals are generally smaller and offer fewer services than urban hospitals.

4) *Teaching Status – teaching or nonteaching*. The missions of teaching hospitals differ from nonteaching hospitals. In addition, financial considerations differ between these two hospital groups. Currently, the Medicare DRG payments are uniformly higher to teaching hospitals than to nonteaching hospitals. A hospital is considered to be a teaching hospital if it has an AMA-approved residency program or is a member of the Council of Teaching Hospitals (COTH).

5) *Bedsize – small, medium, and large.* Bedsize categories are based on hospital beds, and are specific to the hospital's location and teaching status, as shown in Table 6.

Table 6. Bedsize Categories

Location and Teaching Status	**Hospital Bedsize**		
	Small	**Medium**	**Large**
Rural	1-49	50-99	100+
Urban, nonteaching	1-99	100-199	200+
Urban, teaching	1-299	300-499	500+

Rural hospitals were not split according to teaching status, because rural teaching hospitals were rare. For example, in 1988 there were only 20 rural teaching hospitals. The bedsize categories were defined within location and teaching status because they would otherwise have been redundant. Rural hospitals tend to be small; urban nonteaching hospitals tend to be medium-sized; and urban teaching hospitals tend to be large. Yet it was important to recognize gradations of size within these types of hospitals.

For example, in serving rural discharges, the role of "large" rural hospitals (particularly rural referral centers) often differs from the role of "small" rural hospitals. The cut-off points for the bedsize categories are consistent with those used in *Hospital Statistics,* published annually by the AHA.

To further ensure geographic representativeness, implicit stratification variables included state and three-digit zip code (the first three digits of the hospital's five-digit zip code). The hospitals were sorted according to these variables prior to systematic sampling.

HOSPITAL SAMPLING FRAME

For each calendar year, the *universe* of hospitals was established as all community hospitals located in the U.S. However, it was not feasible to obtain and process all-payer discharge data from a random sample of the entire universe of hospitals for at least two reasons. First, all-payer discharge data were not available from all hospitals for research purposes. Second, based on the experience of prior hospital discharge data collections, it would have been too costly to obtain data from individual hospitals, and it would have been too burdensome to process each hospital's unique data structure.

Therefore, the NIS *sampling frame* was constructed from the subset of universe hospitals that released their discharge data for research use. Two sources for all-payer discharge data were state agencies and private data organizations, primarily state hospital associations. Currently, the Agency for Health Care Policy and Research (AHCPR) has agreements with 22 data sources that maintain statewide, all-payer discharge data files to include their data in the HCUP database. However, only 8 states in 1988 and 11 states in 1989-1992 could be included in this first release, as shown in Table 7. Future releases of the NIS will include more states.

Table 7. States in the Frame for the NIS, Release 1

Calendar Years	States in the Frame
1988	California, Colorado, Florida, Iowa, Illinois, Massachusetts, New Jersey, and Washington
1989-1992	Add Arizona, Pennsylvania, and Wisconsin

The Illinois Health Care Cost Containment Council stipulated that no more than 40 percent of Illinois data could be included in the database for any calendar quarter. As a result, approximately 40 percent of the Illinois community hospital universe was randomly selected for the frame each year using the same methodology used to select the NIS hospital sample. That is, Illinois hospitals were stratified on the stratification variables described above, and a systematic random sample of hospitals was drawn for the frame.

Therefore, the list of the entire frame of hospitals was composed of the 40 percent sample of community hospitals for Illinois and all AHA community hospitals in each of the other frame states *that could be matched to the discharge data provided to HCUP*. If an AHA community hospital could not be matched to the discharge data provided by the data source, it was eliminated from the sampling frame (but not from the universe). Unfortunately, only Florida community hospitals are included in the frame for the South region. It is expected that additional southern states will be included in future releases.

The number of frame hospitals for each year is shown in Table 8.

Table 8. Hospital Frame

Calendar Year	Number of Hospitals
1988	1,247
1989	1,658
1990	1,620
1991	1,604
1992	1,591

HOSPITAL SAMPLE DESIGN

Design Requirements

The NIS is a stratified probability sample of hospitals in the frame, with sampling probabilities calculated to select 20 percent of the universe contained in each stratum. The overall objective was to select a sample of hospitals "generalizable" to the target universe, including hospitals outside the frame, which have a zero probability of selection. Moreover, this sample was to be geographically dispersed, yet drawn from the subset of states with inpatient discharge data that agreed to provide such data to the project.

It should be possible, for example, to estimate DRG-specific average lengths of stay over all U.S. hospitals using weighted average lengths of stay, based on averages or regression estimates from the NIS. Ideally, relationships among outcomes and their correlates estimated from the NIS should generally hold across all U.S. hospitals. However, since only 11 states contributed data to this first release, some estimates may be biased. When possible, estimates based on the NIS should be checked against national benchmarks, such as Medicare data or data from the National Hospital Discharge Survey.

The target sample size was 20 percent of the total number of community hospitals in the U.S. for each year in the study period, 1988-1992. This sample size was determined by AHCPR based on their experience with similar research databases.

Alternative stratified sampling allocation schemes were considered. However, allocation proportional to the number of hospitals seemed best for several reasons:

- Fewer than 10 percent of government-planned database applications will produce nationwide estimates. The major government applications will investigate relationships among variables. For example, government researchers will do a substantial amount of regression modeling with these data.

- The HCUP-2 sample[1] used the same stratification and allocation scheme, and it has served AHCPR analysts well. Moreover, the large number of sample hospitals and discharges seemingly reduced the need for variance-reducing allocation schemes.

- AHCPR researchers wanted a simple, easily understood sampling methodology. It was an appealing idea that the NIS sample could be a "miniaturization" of the universe of hospitals (with the obvious geographical limitations imposed by data availability).

- AHCPR statisticians considered other optimal allocation schemes, including sampling hospitals with probabilities proportional to size (number of discharges), and they concluded that sampling with probability proportional to the number of hospitals was preferable. Even though it was recognized that the approach chosen would not be as efficient, the extremely large sample sizes would still yield good estimates. Furthermore, because the data would also be used for purposes other than producing national estimates, it was critical that all hospital types (including small hospitals) be adequately represented.

Hospital Sampling Procedure

Once the universe of hospitals was stratified, up to 20 percent of the total number of U.S. hospitals was randomly selected within each stratum. If too few frame hospitals were in the stratum, then all frame hospitals were selected for the NIS. To simplify variance calculations, at least two hospitals were drawn from each stratum. If fewer than two frame hospitals were contained in a stratum, then that stratum was merged with an "adjacent" stratum containing hospitals with similar characteristics.

We drew a systematic random sample from each stratum, after sorting hospitals by state within each stratum, then by the three-digit zip code (the first three digits of the hospital's five-digit zip code) within each state, and then by a random number within each three-digit zip code. These sorts ensured further geographic generalizability of hospitals within the frame states, and random ordering of hospitals within three-digit zip codes.

Generally, three-digit zip codes that are near in value are geographically near within a state. Furthermore, the U.S. Postal Service locates regional mail distribution centers at the three-digit level. Thus, the boundaries tend to be a compromise between geographic size and population size.

1988 NIS Hospital Sampling Procedure

The 1988 hospital sample was selected according to the following steps:

1. The universe of hospitals was stratified on region, ownership, location, teaching status, and bedsize category.

2. The number of universe and frame hospitals were counted in each stratum.

3. If any stratum had fewer than two hospitals in the *frame,* it was combined with an adjacent stratum to ensure at least two frame hospitals. In all cases where this was required, it was necessary only to collapse the ownership categories. For all cases in which strata were collapsed, private not-for-profit was combined with public, or private investor-owned was combined with public.

4. Within each stratum, the frame hospitals were sorted by state, by three-digit zip code within each state, and by a random number within each three-digit zip code.

5. For each stratum:

   a. The stratum-specific sampling rate (probability) was calculated:

   $$P = \min(1, N/F)$$

   where $N = \max(2, .20*U)$, and U = total number of universe hospitals in the stratum. Therefore, N was the number of hospitals "needed" in each stratum (at least two hospitals and at most 20 percent of the universe).

   F = total number of frame hospitals in the stratum.

   If $F \leq N$ (the number of frame hospitals was less than the number needed), then P = 1 and all hospitals were selected in the stratum.

   b. The skip interval for the systematic sample was calculated:

   $$S = 1 \div P$$

   Every Sth hospital on the list was sampled. For example, if P = .5, then every second hospital was sampled. However, S need *not* have been an integer for this procedure.

   c. A random starting point was calculated for the sample:

   $$R = \text{random number in the interval } (0,S)$$

   d. Every Sth hospital was drawn for the sample from the list of sorted frame hospitals. Let INT(x) be the integer part of x. The first hospital drawn was number INT(1 + R). The second hospital drawn was number INT(1 + R + S). The third hospital drawn was number INT(1 + R + 2S), and so on.

Frame hospitals within a given stratum all had an equal chance of entering the sample. Also, on average, the correct number of hospitals (20 percent of the universe) was drawn for each stratum that had a sufficient number of hospitals.

A total of 758 hospitals was drawn for the 1988 NIS. This number fell short of the overall target of 1,121 hospitals (20 percent of the universe), because several strata contained too few frame hospitals to meet the 20 percent target. More details on the final sample are described later in this report.

1989-1992 NIS Hospital Sampling Procedure

Once the 1988 hospital sample was drawn, it was necessary to draw the 1989 sample by a procedure that "reselected" most of the 1988 hospitals, while allowing hospitals new to the frame an opportunity to enter the 1989 NIS. In particular, hospitals in three states (AZ, PA, and WI) that were not in the 1988 frame entered the 1989 frame.

Even in other frame states, hospitals that opened in 1989 needed a chance to enter the sample. Also, hospitals that changed strata between 1988 and 1989 were considered new to the 1989 frame.

Likewise, once the 1989 hospital sample was drawn, it was necessary to draw the 1990 sample in a way that retained most of the 1989 sample hospitals, while allowing new frame hospitals a chance of selection in 1990.

Consequently, a recursive procedure was developed to update the sample from year to year in a way that properly accounted for changes in stratum size, composition, and sampling rate. The goal of this procedure was to maximize the year-to-year overlap among sample hospitals, yet keep the sampling rate constant for all hospitals *within a stratum*.

The 1988 sampling procedure determined the probability of selection for available frame hospitals within each stratum (probability P). It also gave a procedure for selecting a systematic sample of frame elements with this probability. This procedure was taken as a starting point.

The following procedure provides rules for creating a "year 2" sample, given that a "year 1" sample had already been drawn. For example, year 1 could be 1988 and year 2 could be 1989, or year 1 could be 1989 and year 2 could be 1990. All notation is assumed to refer to sizes and probabilities within a particular stratum.

Probabilities P_1 and P_2 were calculated for sampling hospitals from the frame within the stratum for year 1 and year 2, respectively, based on the frame and universe for year 1 and year 2, respectively. These probabilities were set by the same algorithm used to calculate P for the 1988 hospital sample (step 5a for selecting the 1988 sample).

Now consider the three possibilities associated with changes between years 1 and 2 in the stratum-specific hospital sampling probabilities:

1. $P_2 = P_1$: The target probability was unchanged.

2. $P_2 < P_1$: The target probability decreased.

3. $P_2 > P_1$: The target probability increased.

Below is the procedure used for each of these three cases with one exception: if the stratum-specific probability of selection P_2 was equal to 1, then all frame hospitals were selected for the year 2 sample, regardless of the value of P_1.

Stratum-Specific Sampling Rates the Same ($P_2 = P_1$). If the probability P_2 was the same as P_1, all hospitals in the year 1 sample that remained in the year 2 frame were retained for the year 2 sample. Any new frame hospitals (those in the year 2 frame but not in the year 1 frame) were selected at the rate P_2, using the systematic sampling method described for the 1988 sample selection.

Stratum-Specific Sampling Rate Decreased ($P_2 < P_1$). Now consider the case where the probability of selection decreased between years 1 and 2. First, hospitals new to the frame were sampled with probability P_2. Second, hospitals previously selected for the year 1 sample (that remained in the year 2 frame) were selected for the year 2 sample with probability $P_2 \div P_1$.

The justification for this second procedure was straightforward. For the year 1 sample hospitals that stayed in the frame, the year 1 sample was viewed as the first stage of a two-stage sampling process. The first stage was carried out at the sampling rate of P_1. The second stage was carried out at the sampling rate of $P_2 \div P_1$. Consequently, the "overall" probability of selection was $P_1 \times P_2 \div P_1 = P_2$.

Stratum-Specific Sampling Rate Increased ($P_2 > P_1$). The procedures associated with the case in which the probability of selection was increased between year 1 and year 2 were equally straightforward. First, hospitals new to the frame were sampled with probability P_2. Second, hospitals that were selected in year 1 (that remained in the year 2 frame) were selected for the year 2 sample. Third, hospitals that were in the frame for both years 1 and 2, but not selected for the year 1 sample, were selected for the year 2 sample with probability $(P_2-P_1) \div (1-P_1)$.

The justification for this sampling rate, $(P_2-P_1) \div (1-P_1)$, is somewhat complex. In year 1 certain frame hospitals were included in the sample at the rate P_1. This can also be viewed as having excluded a set of hospitals at the rate $(1-P_1)$. Likewise, in year 2 it was imperative that each hospital excluded from the year 1 sample be excluded from the year 2 sample at an overall rate of $(1-P_2)$.

Since $P_2 > P_1$, then $(1-P_2) < (1-P_1)$. Therefore, just as was done for the case of $P_2 < P_1$, multistage selection was implemented. However, it was implemented for exclusion rather than inclusion.

Therefore, those hospitals excluded from the year 1 sample were also excluded from the year 2 sample at the rate $S = (1-P_2) \div (1-P_1)$. This gave them the desired overall *exclusion* rate of $(1-P_1) \times (1-P_2) \div (1-P_1) = (1-P_2)$. Consequently, the *inclusion* rate for these hospitals was set at $1-S = (P_2-P_1) \div (1-P_1)$.

Zero-Weight Hospitals

To enhance researchers' ability to study the effects of hospital splits and merges, if a hospital was the result of either a split or a merge involving one or more NIS sample hospitals, it was added to the NIS file. However, unless it was selected as a part of the regular NIS sample, it was assigned a sampling weight of zero. Also, any NIS hospital that closed (according to the AHA) was retained in the NIS file and assigned sample weights of zero, if it was not selected for the regular NIS sample in the year it closed. These zero-weight hospitals were included in all following years if inpatient data were available. However, no attempt was made to include these zero-weight hospitals in previous years. For example, if a hospital first appeared in 1990 as a zero-weight hospital, then the hospital would also be added to the 1991-1992 NIS files, but not the 1988-1989 NIS files.

Ten Percent Subsamples

Two non-overlapping 10 percent subsamples of discharges were drawn from the NIS file for each year. The subsamples were selected by drawing every tenth discharge starting with two different starting points (randomly selected between 1 and 10). Having a different starting point for each of the two subsamples guaranteed that they would not overlap. Discharges were sampled so that 10 percent of each hospital's discharges in each quarter were selected for each of the subsamples. The two samples can be combined to form a single, generalizable 20 percent subsample of discharges.

FINAL HOSPITAL SAMPLE

The annual numbers of hospitals and discharges in the NIS, Release 1 are shown in Table 9, for both the regular NIS sample and the total sample (which includes zero-weight hospitals).

Table 9. NIS Hospital Sample

Calendar Year	Regular Sample		Total Sample	
	Number of Hospitals	Number of Discharges	Number of Hospitals	Number of Discharges
1988	758	5,242,904	759	5,265,756
1989	875	6,067,667	882	6,110,064
1990	861	6,156,638	871	6,268,515
1991	847	5,984,270	859	6,156,188
1992	838	6,008,001	856	6,195,744
Total		29,459,480		29,996,267

A more detailed breakdown of the regular NIS hospital sample (excluding zero-weight hospitals), by calendar year and geographic region is shown in Table 10. For each calendar year and each geographic region, Table 10 shows the number of:

- universe hospitals (Universe),
- frame hospitals (Frame),
- sampled hospitals (Sample),
- target hospitals (Target = 20 percent of the universe), and
- shortfall hospitals (Shortfall = Sample - Target).

For example, in 1988 the Northeast region contained 825 hospitals in the universe. It also contained 193 hospitals in the frame, of which 141 hospitals were drawn for the sample. This was 24 hospitals short of the overall target sample size of 165.

From Table 10 it is clear that most of the 1988 shortfall occurred in the North Central and Southern regions. The addition of Wisconsin to the frame in 1989 significantly reduced the shortfall in the North Central region. However, the large shortfall of over 200 hospitals in the Southern region persisted throughout the study period 1988-1992, because only Florida hospitals were in the frame for this release of the NIS.

Table 11 shows the number of hospitals in the universe, frame, and regular sample for each state in the sampling frame for 1988 and 1992. In all states except Illinois, the difference between the universe and the frame represents the number of AHA community hospitals for which no data were received from that state's data source. As explained earlier, the number of hospitals in the Illinois frame is approximately 40 percent of the number in the Illinois universe, as stipulated in agreements with the data source.

The number of hospitals in the NIS hospital sample that continue across multiple sample years is shown in Table 12. From Table 12 it is clear that longitudinal cohorts that include 1988 are the smallest, because the total number of sample hospitals was smallest for 1988 (758 hospitals). However, if 1989 is taken as a starting year, it can then be seen that 93.1 percent of the 1989 hospital sample continued in the 1990 sample (815 of 875). Likewise, the 87.2 percent and 81.0 percent of the 1989 sample hospitals continued on through 1991 and 1992, respectively.

Table 10. Number of Hospitals: Universe, Frame, Regular Sample, Target, and Shortfall By Year and Region

Calendar Year	Region	Universe	Frame	Sample	Target	Shortfall
1988	NE	825	193	141	165	-24
	NC	1,600	208	206	320	-114
	S	2,132	224	200	426	-226
	W	1,050	622	211	210	1
	Total	5,607	1,247	758	1,121	-363
1989	**Region**					
	NE	813	423	165	163	2
	NC	1,582	340	305	316	-11
	S	2,114	222	199	423	-224
	W	1,039	673	206	208	-2
	Total	5,548	1,658	875	1,110	-235
1990	**Region**					
	NE	806	412	166	161	5
	NC	1,574	338	304	315	-11
	S	2,076	218	197	415	-218
	W	1,012	652	194	202	-8
	Total	5,468	1,620	861	1,094	-233
1991	**Region**					
	NE	798	406	162	160	2
	NC	1,560	337	297	312	-15
	S	2,056	215	195	411	-216
	W	998	646	193	200	-7
	Total	5,412	1,604	847	1,082	-235
1992	**Region**					
	NE	790	408	167	158	9
	NC	1,543	334	293	309	-16
	S	2,018	209	192	404	-212
	W	983	640	186	197	-11
	Total	5,334	1,591	838	1,067	-229

Due to rounding, values for regions may not sum to total.

Table 11. Number of Hospitals in the Universe, Frame, and Regular Sample for Each State in the Sampling Frame: 1988 and 1992

Calendar Year	State	Universe	Frame	Sample
1988	CA	471	463	140
	CO	80	62	29
	FL	238	224	200
	IA	127	121	119
	IL	221	87	87
	MA	110	103	83
	NJ	92	90	58
	WA	99	97	42
	Total	1,438	1,247	758
1992	**State**			
	AZ	60	47	15
	CA	437	434	114
	CO	71	69	29
	FL	224	209	192
	IA	121	119	106
	IL	211	87	79
	MA	102	92	43
	NJ	97	89	32
	PA	232	227	92
	WA	91	90	28
	WI	128	128	108
	Total	1,774	1,591	838

Table 12. Number of Hospitals and Discharges in Longitudinal Cohort

Number of Years	Calendar Years	Longitudinal Regular Sample Hospitals	% of Base Year Sample	Longitudinal Regular Sample Discharges
2	1988-1989	610	80.5	8,492,039
	1989-1990	815	93.1	11,525,749
	1990-1991	802	93.1	11,297,175
	1991-1992	781	92.2	11,272,981
3	1988-1990	573	75.6	12,168,677
	1989-1991	763	87.2	16,074,381
	1990-1992	745	86.5	16,085,651
4	1988-1991	542	71.5	15,096,807
	1989-1992	709	81.0	20,340,970
5	1988-1992	502	66.2	18,106,098

SAMPLING WEIGHTS

Although the sampling design was simple and straightforward, it is necessary to incorporate sample weights to obtain state and national estimates. Therefore, sample weights were developed separately for hospital- and discharge-level analyses for each year from 1988 to 1992. Three hospital-level weights were developed to weight NIS sample hospitals to the state, frame, and universe. Similarly, three discharge-level weights were developed to weight NIS sample discharges to the state, frame, and universe.

Hospital-Level Sampling Weights

Universe Hospital Weights. Hospital weights to the universe were calculated by post-stratification. For each calendar year, hospitals were stratified on the same variables that were used for sampling: geographic region, urban/rural location, teaching status, bedsize, and ownership. The strata that were collapsed for sampling were also collapsed for sample weight calculations. Within stratum s, each NIS sample hospital's universe weight was calculated as:

$$W_s(\text{universe}) = N_s(\text{universe}) \div N_s(\text{sample}),$$

where N_s(universe) and N_s(sample) were the number of community hospitals within stratum s in the universe and sample, respectively. Thus, each hospital's universe weight is equal to the number of universe hospitals it represented during that calendar year.

Frame Hospital Weights. Hospital-level sampling weights were also calculated to represent the entire collection of states in the frame using the same post-stratification scheme as described above for the weights to represent the universe. For each year, within stratum s, each NIS sample hospital's frame weight was calculated as:

$$W_s(\text{frame}) = N_s(\text{frame}) \div N_s(\text{sample}).$$

N_s(frame) was the total number of universe community hospitals within stratum s in the states that contributed data to the frame. N_s(sample) was the number of sample hospitals selected for the NIS in stratum s. Thus, each hospital's frame weight is equal to the number of universe hospitals it represented in the frame states during that calendar year.

State Hospital Weights. For each year, a hospital's weight to its state was calculated in a similar fashion. Within each state, strata often had to be collapsed after sample selection for development of weights to ensure a minimum of two sample hospitals within each stratum. For each state and each year, within stratum s, each NIS sample hospital's state weight was calculated as:

$$W_s(\text{state}) = N_s(\text{state}) \div N_s(\text{state sample}).$$

N_s(state) was the number of universe community hospitals in the state within stratum s. N_s(state sample) was the number of hospitals selected for the NIS from that state in stratum s. Thus, each hospital's state weight is equal to the number of hospitals that it represented in its state during that calendar year.

All of these hospital weights can be rescaled if necessary for selected analyses, to sum to the NIS hospital sample size each year.

Discharge-Level Sampling Weights

The calculations for discharge-level sampling weights were very similar to the calculations of hospital-level sampling weights. The discharge weights usually are constant for all discharges within a stratum.

The only exceptions were for strata with sample hospitals that, according to the AHA files, were open for the entire calendar year but contributed less than their full year of data to the NIS. For those hospitals, we *adjusted* the number of observed discharges by a factor 4 ÷ Q, where Q was the number of calendar quarters that the hospital contributed discharges to the NIS. For example, when a sample hospital contributed only two quarters of discharge data to the NIS, the *adjusted* number of discharges was double the observed number.

With that minor adjustment, each discharge weight is essentially equal to the number of reference (universe, frame, or state) discharges that each sampled discharge represented in its stratum. This calculation was possible because the number of total discharges was available for every hospital in the universe from the AHA files. Each universe hospital's AHA discharge total was calculated as the sum of newborns and total facility discharges.

Universe Discharge Weights. Discharge weights to the universe were calculated by post-stratification. For each calendar year, hospitals were stratified just as they were for universe hospital weight calculations. Within stratum s, for hospital i, each NIS sample discharge's universe weight was calculated as:

$$DW_{is}(\text{universe}) = [DN_s(\text{universe}) \div ADN_s(\text{sample})] * (4 \div Q_i),$$

where DN_s(universe) was the number of discharges from community hospitals in the universe within stratum s; ADN_s(sample) was the number of *adjusted* discharges from sample hospitals selected for the NIS; and Q_i was the number of quarters of discharge data contributed by hospital i to the NIS (usually $Q_i = 4$). Thus, each discharge's weight is equal to the number of universe discharges it represented in stratum s during that calendar year.

Frame Discharge Weights. Discharge-level sampling weights were also calculated to represent all discharges from the entire collection of states in the frame using the same post-stratification scheme described above for the discharge weights to represent the universe. For each year, within stratum s, for hospital i, each NIS sample discharge's frame weight was calculated as:

$$W_{is}(\text{frame}) = [DN_s(\text{frame}) \div ADN_s(\text{sample})] * (4 \div Q_i),$$

DN_s(frame) was the number of discharges from all community hospitals in the states that contributed to the frame within stratum s. ADN_s(sample) was the number of *adjusted* discharges from sample hospitals selected for the NIS in stratum s. Q_i was the number of quarters of discharge data contributed by hospital i to the NIS (usually $Q_i = 4$). Thus, each discharges's frame weight is equal to the number of discharges it represented in the frame states during that calendar year.

State Discharge Weights. For each year, a discharge's weight to its state was similarly calculated. Strata were collapsed in the same way as they were for the state hospital weights to ensure a minimum of two sample hospitals within each stratum. For each year, within stratum s, for hospital i, each NIS sample discharge's state weight was calculated as:

$$W_{is}(\text{state}) = [DN_s(\text{state}) \div ADN_s(\text{state sample})] * (4 \div Q_i),$$

DN_s(state) was the number of discharges from all community hospitals in the state within stratum s. ADN_s(state sample) was the *adjusted* number of discharges from hospitals selected for the NIS from that state in stratum s. Q_i was the number of quarters of discharge data contributed by hospital i to the NIS (usually $Q_i = 4$). Thus, each discharge's state weight is equal to the number of discharges that it represented in its state during that calendar year.

All of these discharge weights can be rescaled if necessary for selected analyses, to sum to the NIS discharge sample size each year.

Discharge Weights for 10 Percent Subsamples

In the 10 percent subsamples, each discharge had a 10 percent chance of being drawn. Therefore, the discharge weights contained in the Hospital Weights file can be multiplied by 10 for each of the subsamples, or multiplied by 5 for the two subsamples combined.

DATA ANALYSIS

Variance Calculations

It may be important for researchers to calculate a measure of precision for some estimates based on the NIS sample data. Variance estimates must take into account both the sampling design and the form of the statistic. The sampling design was a stratified, single-stage cluster sample. A stratified random sample of hospitals (clusters) was drawn and then *all* discharges were included from each selected hospital.

If hospitals inside the frame were similar to hospitals outside the frame, the sample hospitals can be treated as if they were randomly selected from the entire universe of hospitals within each stratum. Standard formulas for a stratified, single-stage cluster sampling without replacement could be used to calculate statistics and their variances in most applications.

A multitude of statistics can be estimated from the NIS data. Several computer programs are listed below that calculate statistics and their variances from sample survey data. Some of these programs use general methods of variance calculations (e.g., the jackknife and balanced half-sample replications) that take into account the sampling design. However, it may be desirable to calculate variances using formulas specifically developed for some statistics.

In most cases, computer programs are readily available to perform these calculations. For instance, OSIRIS IV, developed at the University of Michigan, does calculations for numerous statistics arising from the stratified, single-stage cluster sampling design.

These variance calculations are based on finite-sample theory, which is an appropriate method for obtaining cross-sectional, nationwide estimates of outcomes. According to finite-sample theory, the intent of the estimation process is to obtain estimates that are precise representations of the nationwide population at a specific point in time. In the context of the NIS, any estimates that attempt to accurately describe characteristics (such as expenditure and utilization patterns or hospital market factors) and interrelationships among characteristics of hospitals and discharges during a specific year from 1988 to 1992 should be governed by finite-sample theory.

Alternatively, in the study of hypothetical population outcomes not limited to a specific point in time, analysts may be less interested in specific characteristics from the finite population (and time period) from which the *sample* was drawn, than they are in hypothetical characteristics of a conceptual "superpopulation" from which any particular finite *population* in a given year might have been drawn. According to this superpopulation model, the nationwide population in a given year is only a snapshot in time of the possible interrelationships among hospital, market, and discharge characteristics. In a given year, all possible interactions between such characteristics may not have been observed, but analysts may wish to predict or simulate interrelationships that may occur in the future.

Under the finite-population model, the variances of estimates approach zero as the sampling fraction approaches one, since the population is defined at that point in time, and because the estimate is for a characteristic as it existed at the time of sampling. This is in contrast to the superpopulation model, which adopts a stochastic viewpoint rather than a deterministic viewpoint. That is, the nationwide population in a particular year is viewed as a random sample of some underlying superpopulation over time.

Different methods are used for calculating variances under the two sample theories. Under the superpopulation (stochastic) model, procedures (such as those described by Potthoff, Woodbury, and Manton[2]) have been developed to draw inferences using weights from complex samples. In this context, the survey weights are not used to weight the sampled cases to the universe, because the universe is conceptually infinite in size. Instead, these weights are used to produce unbiased estimates of parameters that govern the superpopulation.

In summary, the choice of an appropriate method for calculating variances for nationwide estimates depends on the type of measure and the intent of the estimation process.

Computer Software for Variance Calculations

The hospital weights will be useful for producing hospital-level statistics for analyses that use the *hospital* as the unit of analysis, and the discharge weights will be useful for producing discharge-level statistics for analyses that use the *discharge* as the unit of analysis. These would be used to weight the sample data in estimating population statistics.

Several statistical programming packages allow weighted analyses.[3] For example, nearly all SAS (Statistical Analysis System) procedures incorporate weights.

In addition, several publicly available subroutines have been developed specifically for calculating statistics and their standard errors from survey data:

- OSIRIS IV was developed by L. Kish, N. Van Eck, and M. Frankel at the Survey Research Center, University of Michigan. It consists of two main programs for estimating variances from complex survey designs.

- SUDAAN, a set of SAS subroutines, was developed at the Research Triangle Institute by B. V. Shah. It is adequate for handling most survey designs with stratification. The procedures can handle estimation and variance estimation for means, proportions, ratios, and regression coefficients.

- SUPER CARP (Cluster Analysis and Regression Program) was developed at Iowa State University by W. Fuller, M. Hidiroglou, and R. Hickman. This program computes estimates and variance estimates for multistage, stratified sampling designs with arbitrary probabilities of selection. It can handle estimated totals, means, ratios, and regression estimates.

The NIS database includes a Hospital Weights file with variables required by these programs to calculate finite population statistics. In addition to the sample weights described earlier, hospital identifiers (PSUs), stratification variables, and stratum-specific totals for the numbers of discharges and hospitals are included so that finite-population corrections (FPCs) can be applied to variance estimates.

In addition to these subroutines, standard errors can be estimated by validation and cross-validation techniques. Given that a very large number of observations will be available for most analyses, it may be feasible to set aside a part of the data for validation purposes. Standard errors and confidence intervals can then be calculated from the validation data. If the analytical file is too small to set aside a large validation sample, cross-validation techniques may be used.

For example, tenfold cross-validation would split the data into ten equal-sized subsets. The estimation would take place in ten iterations. At each iteration, the outcome of interest is predicted for one-tenth of the observations by an estimate based on a model fit to the other nine-tenths of the observations. Unbiased estimates of error variance are then obtained by comparing the actual values to the predicted values obtained in this manner.

Finally, it should be noted that a large array of hospital-level variables are available for the entire universe of hospitals, including those outside the sampling frame. For instance, the variables from the AHA surveys and from the Medicare Cost Reports are available for nearly all hospitals. To the extent that hospital-level outcomes correlate with these variables, they may be used to sharpen regional and nationwide estimates.

As a simple example, each hospital's number of C-sections would be correlated with their total number of deliveries. The number of C-sections must be obtained from discharge data, but the number of deliveries is available from AHA data. Thus, if a regression can be fit predicting C-sections from deliveries based on the NIS data, that regression can then be used to obtain hospital-specific estimates of the number of C-sections for all hospitals in the universe.

Longitudinal Analyses

As previously shown in Table 12, hospitals that continue in the NIS for multiple consecutive years are a subset of the hospitals in the NIS for any one of those years. Consequently, longitudinal analyses of hospital-level outcomes may be biased if they are based on any subset of NIS hospitals limited to continuous NIS membership. In particular, such subsets would tend to contain fewer hospitals that opened, closed, split, merged, or changed strata. Further, the sample weights were developed as annual, cross-sectional weights rather than longitudinal weights. Therefore, different weights might be required, depending on the statistical methods employed by the analyst.

One approach to consider in hospital-level longitudinal analyses is to use repeated-measure models that allow hospitals to have missing values for some years. However, the data are not actually missing for some hospitals, such as those that closed during the study period. In any case, the analyses may be more efficient (e.g., produce more precise estimates) if they account for the potential correlation between repeated measures on the same hospital over time, yet incorporate data from all hospitals in the sample during the study period.

Discharge Subsamples

The two non-overlapping 10 percent subsamples of discharges were drawn from the NIS file for each year for several reasons pertaining to data analysis. One reason for creating the subsamples was to reduce processing costs for selected studies that will not require the entire NIS. Another reason is that the two subsamples may be used to validate models and obtain unbiased estimates of standard errors. That is, one subsample may be used to estimate statistical models, and the other subsample may be used to test the fit of those models on new data. This is a very important analytical step, particularly in exploratory studies, where one runs the risk of fitting noise.

For example, it is well known that the percentage of variance explained by a regression, R^2, is generally overestimated by the data used to fit a model. The regression model could be estimated from the first subsample and then applied to the second subsample. The squared correlation between the actual and predicted value in the second subsample is an unbiased estimate of the model's true explanatory power when applied to new data.

ENDNOTES

1. Coffey, R. and D. Farley (1988, July). *HCUP-2 Project Overview,* (DHHS Publication No. (PHS) 88-3428. Hospital Studies Program Research Note 10, National Center for Health Services Research and Health Care Technology Assessment, Rockville, MD: Public Health Service.

2. Potthoff, R.F., M.A. Woodbury, and K.G. Manton (1992). "Equivalent Sample Size" and "Equivalent Degrees of Freedom" Refinements for Inference Using Survey Weights Under Superpopulation Models. *Journal of the American Statistical Association*, Vol. 87, 383-396.

3. Carlson, B.L., A.E. Johnson, and S.B. Cohen (1993). An Evaluation of the Use of Personal Computers for Variance Estimation with Complex Survey Data. *Journal of Official Statistics*, Vol. 9, No. 4, 795-814.

TECHNICAL SUPPLEMENT 6:
DESIGN OF THE HCUP NATIONWIDE INPATIENT SAMPLE, RELEASE 2

INTRODUCTION

The Nationwide Inpatient Sample (NIS) of the Healthcare Cost and Utilization Project (HCUP) was established to provide analyses of hospital utilization across the United States. The NIS, Release 1 covers calendar years 1988-1992. Release 2 covers calendar year 1993. The target universe includes all acute-care discharges from all community hospitals in the United States; the NIS comprises all discharges from a sample of hospitals in this target universe.

This second release of the NIS contains 6.5 million discharges from a sample of 913 hospitals in 17 states. The first release (1988 through 1992) contains 5.2 to 6.2 million discharges per year from a sample of 758 to 875 hospitals per year in 11 states (8 states for 1988). Thus, the NIS supports both cross-sectional and longitudinal analyses.

Potential research issues focus on both discharge- and hospital-level outcomes. Discharge outcomes of interest include trends in inpatient treatments with respect to:

- frequency,
- costs,
- lengths of stay,
- effectiveness,
- appropriateness, and
- access to hospital care.

Hospital outcomes of interest include:

- mortality rates,
- complication rates,
- patterns of care,
- diffusion of technology, and
- trends toward specialization.

These and other outcomes are of interest for the nation as a whole and for policy-relevant inpatient subgroups defined by geographic regions, patient demographics, hospital characteristics, physician characteristics, and pay sources.

This report provides a detailed description of the NIS, Release 2 sample design, as well as a summary of the resultant hospital sample. Sample weights were developed to obtain national estimates of hospital and inpatient parameters. These weights and other special-use weights are described in detail. Tables include cumulative information for NIS, Release 1 (1988 through 1992) and NIS, Release 2 (1993) to provide a longitudinal view of the database.

THE NIS HOSPITAL UNIVERSE

The hospital universe is defined by all hospitals that were open during any part of the calendar year and were designated as community hospitals in the American Hospital Association (AHA) Annual Survey of Hospitals. For purposes of the NIS, the definition of a community hospital is that used by the AHA: "all nonfederal short-term general and other specialty hospitals, excluding hospital units of institutions." Consequently, Veterans Hospitals and other federal hospitals are excluded.

Table 13 shows the number of universe hospitals for each year based on the AHA Annual Survey.

Table 13. Hospital Universe

Year	Number of Hospitals
1988	5,607
1989	5,548
1990	5,468
1991	5,412
1992	5,334
1993	5,313

Hospital Merges, Splits, and Closures

All hospital entities that were designated community hospitals in the AHA hospital file were included in the hospital universe. Therefore, if two or more community hospitals merged to create a new community hospital, the original hospitals and the newly-formed hospital were all considered separate hospital entities in the universe for the year of the merge. Likewise, if a community hospital split, the original hospital and all newly created community hospitals were separate entities in the universe for the year of the split. Finally, community hospitals that closed during a year were included as long as they were in operation during some part of the calendar year.

Stratification Variables

To help ensure representativeness, sampling strata were defined based on five hospital characteristics contained in the AHA hospital files. The stratification variables were as follows:

1) *Geographic Region – Northeast, Midwest, West, and South*. This is an important stratifier because practice patterns have been shown to vary substantially by region. For example, lengths of stay tend to be longer in East Coast hospitals than in West Coast hospitals.

2) *Control – government nonfederal, private not-for-profit, and private investor-owned.* These types of hospitals tend to have different missions and different responses to government regulations and policies.

3) *Location – urban or rural.* Government payment policies often differ according to this designation. Also, rural hospitals are generally smaller and offer fewer services than urban hospitals.

4) *Teaching Status – teaching or nonteaching*. The missions of teaching hospitals differ from nonteaching hospitals. In addition, financial considerations differ between these two hospital groups. Currently, the Medicare DRG payments are uniformly higher to teaching hospitals than to nonteaching hospitals. A hospital is considered to be a teaching hospital if it has an AMA-approved residency program or is a member of the Council of Teaching Hospitals (COTH).

5) *Bedsize – small, medium, and large.* Bedsize categories are based on hospital beds, and are specific to the hospital's location and teaching status, as shown in Table 14.

Table 14. Bedsize Categories

Location and Teaching Status	Hospital Bedsize		
	Small	Medium	Large
Rural	1-49	50-99	100+
Urban, nonteaching	1-99	100-199	200+
Urban, teaching	1-299	300-499	500+

Rural hospitals were not split according to teaching status, because rural teaching hospitals were rare. For example, in 1988 there were only 20 rural teaching hospitals. The bedsize categories were defined within location and teaching status because they would otherwise have been redundant. Rural hospitals tend to be small; urban nonteaching hospitals tend to be medium-sized; and urban teaching hospitals tend to be large. Yet it was important to recognize gradations of size within these types of hospitals.

For example, in serving rural discharges, the role of "large" rural hospitals (particularly rural referral centers) often differs from the role of "small" rural hospitals. The cut-off points for the bedsize categories are consistent with those used in *Hospital Statistics,* published annually by the AHA.

To further ensure geographic representativeness, implicit stratification variables included state and three-digit zip code (the first three digits of the hospital's five-digit zip code). The hospitals were sorted according to these variables prior to systematic sampling.

HOSPITAL SAMPLING FRAME

For each year, the *universe* of hospitals was established as all community hospitals located in the U.S. However, it was not feasible to obtain and process all-payer discharge data from a random sample of the entire universe of hospitals for at least two reasons. First, all-payer discharge data were not available from all hospitals for research purposes. Second, based on the experience of prior hospital discharge data collections, it would have been too costly to obtain data from individual hospitals, and it would have been too burdensome to process each hospital's unique data structure.

Therefore, the NIS *sampling frame* was constructed from the subset of universe hospitals that released their discharge data for research use. Two sources for all-payer discharge data were state agencies and private data organizations, primarily state hospital associations. At the time when the sample was drawn, the Agency for Health Care Policy and Research (AHCPR) had agreements with 22 data sources that maintain statewide, all-payer discharge data files to include their data in the HCUP database. However, only 8 states in 1988 and 11 states in 1989-1992 could be included in the first release of the NIS, and an additional 6 states have been included in the second release of the NIS, as shown in Table 15.

Table 15. States in the Frame for the NIS, Release 1 and NIS, Release 2

Years	States in the Frame
NIS, Release 1	
1988	California, Colorado, Florida, Iowa, Illinois, Massachusetts, New Jersey, and Washington
1989-1992	Add Arizona, Pennsylvania, and Wisconsin
NIS, Release 2	
1993	Add Connecticut, Kansas, Maryland, New York, Oregon, South Carolina

The list of the entire frame of hospitals was composed of all AHA community hospitals in each of the frame states *that could be matched to the discharge data provided to HCUP*, with restrictions on the hospitals that could be included from Illinois and South Carolina. If an AHA community hospital could not be matched to the discharge data provided by the data source, it was eliminated from the sampling frame (but not from the universe).

The Illinois Health Care Cost Containment Council stipulated that no more than 40 percent of the data provided by Illinois could be included in the database for any calendar quarter. As a result, the number of Illinois community hospitals in the frame was restricted, and 105 of the 209 Illinois community hospital universe (50 percent of hospitals) were randomly selected using the same methodology used to select the NIS hospital sample. That is, Illinois hospitals were stratified on the stratification variables described above, and a systematic random sample of hospitals was drawn for the frame. This prevented the sample from including more than 40 percent of Illinois discharges.

South Carolina stipulated that only hospitals that appear in sampling strata with two or more hospitals were to be included in the NIS. Five South Carolina hospitals were excluded from the frame since there were fewer than two South Carolina hospitals in five sampling frame strata. The remaining 60 South Carolina community hospitals are included in the frame.

The number of frame hospitals for each year is shown in Table 16.

Table 16. Hospital Frame

Year	Number of Hospitals
1988	1,247
1989	1,658
1990	1,620
1991	1,604
1992	1,591
1993	2,168

HOSPITAL SAMPLE DESIGN

Design Requirements

The NIS is a stratified probability sample of hospitals in the frame, with sampling probabilities calculated to select 20 percent of the universe contained in each stratum. The overall objective was to select a sample of hospitals "generalizable" to the target universe, which includes hospitals outside the frame (zero probability of selection). Moreover, this sample was to be geographically dispersed, yet drawn from the subset of states with inpatient discharge data that agreed to provide such data to the project.

It should be possible, for example, to estimate DRG-specific average lengths of stay over all U.S. hospitals using weighted average lengths of stay, based on averages or regression estimates from the NIS. Ideally, relationships among outcomes and their correlates estimated from the NIS should generally hold across all U.S. hospitals. However, since only 17 states contributed data to this second release, some estimates may differ from estimates from comparative data sources. When possible, estimates based on the NIS should be checked against national benchmarks, such as Medicare data or data from the National Hospital Discharge Survey to determine the appropriateness of the NIS for specific analyses.

The target sample size was 20 percent of the total number of community hospitals in the U.S. for 1993. This sample size was determined by AHCPR based on their experience with similar research databases.

Alternative stratified sampling allocation schemes were considered. However, allocation proportional to the number of hospitals is preferred for several reasons:

- Fewer than 10 percent of government-planned database applications will produce nationwide estimates. The major government applications will investigate relationships among variables. For example, government researchers will do a substantial amount of regression modeling with these data.

- The HCUP-2 sample[1] used the same stratification and allocation scheme, and it has served AHCPR analysts well. Moreover, the large number of sample hospitals and discharges seemingly reduced the need for variance-reducing allocation schemes.

- AHCPR researchers wanted a simple, easily understood sampling methodology. It was an appealing idea that the NIS sample could be a "miniaturization" of the universe of hospitals (with the obvious geographical limitations imposed by data availability).

- AHCPR statisticians considered other optimal allocation schemes, including sampling hospitals with probabilities proportional to size (number of discharges), and they concluded that sampling with probability proportional to the number of hospitals was preferable. Even though it was recognized that the approach chosen would not be as efficient, the extremely large sample sizes yield good estimates. Furthermore, because the data are to be used for purposes other than producing national estimates, it is critical that all hospital types (including small hospitals) are adequately represented.

Hospital Sampling Procedure

Once the universe of hospitals was stratified, up to 20 percent of the total number of U.S. hospitals was randomly selected within each stratum. If too few frame hospitals were in the stratum, then all frame hospitals were selected for the NIS, subject to sampling restrictions specified by states. To simplify variance calculations, at least two hospitals were drawn from each stratum. If fewer than two frame hospitals were contained in a stratum, then that stratum was merged with an "adjacent" stratum containing hospitals with similar characteristics.

A systematic random sample was drawn from each stratum, after sorting hospitals by state within each stratum, then by the three-digit zip code (the first three digits of the hospital's five-digit zip code) within each state, and then by a random number within each three-digit zip code. These sorts ensured further geographic generalizability of hospitals within the frame states, and random ordering of hospitals within three-digit zip codes.

Generally, three-digit zip codes that are near in value are geographically near within a state. Furthermore, the U.S. Postal Service locates regional mail distribution centers at the three-digit level. Thus, the boundaries tend to be a compromise between geographic size and population size.

1993 NIS Hospital Sampling Procedures

The 1993 sample was drawn by a procedure that retained most of the 1992 hospitals, while allowing hospitals new to the frame an opportunity to enter the 1993 NIS. In particular, hospitals in six states (CT, KS, MD, NY, OR, and SC) that were not in the 1992 frame entered the 1993 frame.

Even in frame states that were present in the 1992 sample, hospitals that opened in 1993 needed a chance to enter the sample. Also, hospitals that changed strata between 1992 and 1993 were considered new to the 1993 frame.

Consequently, a recursive procedure was developed to update the sample from year to year in a way that properly accounted for changes in stratum size, composition, and sampling rate. The goal of this procedure was to maximize the year-to-year overlap among sample hospitals, yet keep the sampling rate constant for all hospitals *within a stratum*.

The following procedure provides rules for creating a "year 2" sample, given that a "year 1" sample had already been drawn. In this example, year 1 would be 1992 and year 2 would be 1993. All notation is assumed to refer to sizes and probabilities within a particular stratum.

Probabilities P_1 and P_2 were calculated for sampling hospitals from the frame within the stratum for year 1 and year 2, respectively, based on the frame and universe for year 1 and year 2, respectively. These probabilities were set by the same algorithm used to calculate P for the 1988 hospital sample (see Technical Supplement: *Design of the HCUP Nationwide Inpatient Sample, Release 1*, in section "1988 NIS Hospital Sampling Procedure.")

Now consider the three possibilities associated with changes between years 1 and 2 in the stratum-specific hospital sampling probabilities:

1. $P_2 = P_1$: The target probability was unchanged.

2. $P_2 < P_1$: The target probability decreased.

3. $P_2 > P_1$: The target probability increased.

Below is the procedure used for each of these three cases with one exception: if the stratum-specific probability of selection P_2 was equal to 1, then all frame hospitals were selected for the year 2 sample, regardless of the value of P_1.

Stratum-Specific Sampling Rates the Same ($P_2 = P_1$). If the probability P_2 was the same as P_1, all hospitals in the year 1 sample that remained in the year 2 frame were retained for the year 2 sample. Any new frame hospitals (those in the year 2 frame but not in the year 1 frame) were selected at the rate P_2, using the systematic sampling method described for the 1988 sample selection in Technical Supplement: *Design of the HCUP Nationwide Inpatient Sample, Release 1*.

Stratum-Specific Sampling Rate Decreased ($P_2 < P_1$). Now consider the case where the probability of selection decreased between years 1 and 2. First, hospitals new to the frame were sampled with probability P_2. Second, hospitals previously selected for the year 1 sample (that remained in the year 2 frame) were selected for the year 2 sample with probability $P_2 \div P_1$.

The justification for this second procedure was straightforward. For the year 1 sample hospitals that stayed in the frame, the year 1 sample was viewed as the first stage of a two-stage sampling process. The first stage was carried out at the sampling rate of P_1. The second stage was carried out at the sampling rate of $P_2 \div P_1$. Consequently, the "overall" probability of selection was P_1 x $P_2 \div P_1 = P_2$.

Stratum-Specific Sampling Rate Increased ($P_2 > P_1$). The procedures associated with the case in which the probability of selection was increased between year 1 and year 2 were equally straightforward. First, hospitals new to the frame were sampled with probability P_2. Second, hospitals that were selected in year 1 (that remained in the year 2 frame) were selected for the year 2 sample. Third, hospitals that were in the frame for both years 1 and 2, but not selected for the year 1 sample, were selected for the year 2 sample with probability $(P2-P_1) \div (1-P_1)$.

The justification for this sampling rate, $(P_2-P_1) \div (1-P_1)$, is somewhat complex. In year 1 certain frame hospitals were included in the sample at the rate P_1. This can also be viewed as having excluded a set of hospitals at the rate $(1-P_1)$. Likewise, in year 2 it was imperative that each hospital excluded from the year 1 sample be excluded from the year 2 sample at an overall rate of $(1-P_2)$.

Since $P_2 > P_1$, then $(1-P_2) < (1-P_1)$. Therefore, just as was done for the case of $P_2 < P_1$, multistage selection was implemented. However, it was implemented for exclusion rather than inclusion.

Therefore, those hospitals excluded from the year 1 sample were also excluded from the year 2 sample at the rate $S = (1-P_2) \div (1-P_1)$. This gave them the desired overall *exclusion* rate of $(1-P_1) \times (1-P_2) \div (1-P_1) = (1-P_2)$. Consequently, the *inclusion* rate for these hospitals was set at $1-S = (P_2-P_1) \div (1-P_1)$.

Zero-Weight Hospitals

The 1993 sample contains no zero-weight hospitals. For a description of zero-weight hospitals in the 1988-1992 sample, see Technical Supplement: *Design of the HCUP Nationwide Inpatient Sample, Release 1*.

Ten Percent Subsamples

Two nonoverlapping 10 percent subsamples of discharges were drawn from the NIS file for each year. The subsamples were selected by drawing every tenth discharge starting with two different starting points (randomly selected between 1 and 10). Having a different starting point for each of the two subsamples guaranteed that they would not overlap. Discharges were sampled so that 10 percent of each hospital's discharges in each quarter were selected for each of the subsamples. The two samples can be combined to form a single, generalizable 20 percent subsample of discharges.

FINAL HOSPITAL SAMPLE

The annual numbers of hospitals and discharges in NIS, Release 1 and NIS, Release 2 are shown in Table 17, for both the regular NIS sample and the total sample (which includes zero-weight hospitals for 1988-1992).

Table 17. NIS Sample

	Regular Sample		Total Sample	
Year	**Number of Hospitals**	**Number of Discharges**	**Number of Hospitals**	**Number of Discharges**
NIS, Release 1				
1988	758	5,242,904	759	5,265,756
1989	875	6,067,667	882	6,110,064
1990	861	6,156,638	871	6,268,515
1991	847	5,984,270	859	6,156,188
1992	838	6,008,001	856	6,195,744
NIS, Release 2				
1993	913	6,538,976	913	6,538,976
Total		35,998,456		36,535,243

A more detailed breakdown of the 1993 NIS hospital sample by geographic region is shown in Table 18. For each geographic region, Table 18 shows the number of:

- universe hospitals (Universe),
- frame hospitals (Frame),
- sampled hospitals (Sample),
- target hospitals (Target = 20 percent of the universe), and
- shortfall hospitals (Shortfall = Sample - Target).

Table 18. Number of Hospitals in Universe, Frame, Regular Sample, Target, and Shortfall By Region, 1993

Region	Universe	Frame	Sample	Target	Shortfall
NE	789	672	174	158	16
MW	1,533	478	302	307	-5
S	2,013	320	258	403	-145
W	978	698	179	196	-17
Total	5,313	2,168	913	1,064	-151

For example, in 1993 the Northeast region contained 789 hospitals in the universe. It also contained 672 hospitals in the frame, of which 174 hospitals were drawn for the sample. This was 16 hospitals more than the target sample size of 158.

Table 19 shows the number of hospitals in the universe, frame, and regular sample for each state in the sampling frame for 1993. In all states except Illinois and South Carolina, the difference between the universe and the frame represents the number of AHA community hospitals for which no data were received from that state's data source. As explained earlier, the number of hospitals in the Illinois frame is approximately 40 percent of the number in the Illinois universe, as stipulated in agreements with the data source. The number of hospitals in the South Carolina frame is five fewer than the number in the South Carolina universe, as stipulated in agreements with the data source.

Table 19. Number of Hospitals in the Universe, Frame, and Regular Sample for Each State in the Sampling Frame: 1993

State	Universe	Frame	Sample
AZ	60	47	13
CA	431	429	96
CO	72	71	28
CT	35	33	7
FL	224	210	166
IA	119	119	70
IL	209	105	75
KS	136	126	72
MA	99	89	30
MD	50	50	40
NJ	97	89	20
NY	232	231	60
OR	63	61	19
PA	233	230	57
SC	68	60	52
WA	92	90	23
WI	128	128	85
Total	2348	2168	913

The number of hospitals in the NIS hospital samples that continue across multiple sample years is shown in Table 20. This table is of interest only to those who want to combine Release 1 and 2 of the NIS. From Table 20 it is clear that longitudinal cohorts that include 1988 and 1993 are the smallest, because the total number of sample hospitals was smallest for 1988 (758 hospitals) and the sampling frame changed the most in 1993. As an example, if 1989 is taken as a starting year, it can then be seen that 59.8 percent of the 1989 hospital sample continued in the 1993 sample (523 of 875).

Table 20. Number of Hospitals and Discharges in Longitudinal Cohort

Number of Years	Calendar Years	Longitudinal Regular Sample Hospitals	% of Base Year Sample	Longitudinal Regular Sample Discharges
2	1988-1989	610	80.5	8,492,039
	1989-1990	815	93.1	11,525,749
	1990-1991	802	93.1	11,297,175
	1991-1992	781	92.2	11,272,981
	1992-1993	609	72.7	8,804,638
3	1988-1990	573	75.6	12,168,677
	1989-1991	763	87.2	16,074,381
	1990-1992	745	86.5	16,085,651
	1991-1993	570	67.3	12,559,421
4	1988-1991	542	71.5	15,096,807
	1989-1992	709	81.0	20,340,970
	1990-1993	548	63.6	16,023,500
5	1988-1992	502	66.2	18,106,098
	1989-1993	523	59.8	19,000,777
6	1988-1993	378	49.9	16,906,818

SAMPLING WEIGHTS

Although the sampling design was simple and straightforward, it is necessary to incorporate sample weights to obtain state and national estimates. Therefore, sample weights were developed separately for hospital- and discharge-level analyses. Three hospital-level weights were developed to weight NIS sample hospitals to the state, frame, and universe. Similarly, three discharge-level weights were developed to weight NIS sample discharges to the state, frame, and universe.

Hospital-Level Sampling Weights

Universe Hospital Weights. Hospital weights to the universe were calculated by post-stratification. For each year, hospitals were stratified on the same variables that were used for sampling: geographic region, urban/rural location, teaching status, bedsize, and control. The strata that were collapsed for sampling were also collapsed for sample weight calculations. Within stratum s, each NIS sample hospital's universe weight was calculated as:

$$W_s(\text{universe}) = N_s(\text{universe}) \div N_s(\text{sample}),$$

where N_s(universe) and N_s(sample) were the number of community hospitals within stratum s in the universe and sample, respectively. Thus, each hospital's universe weight is equal to the number of universe hospitals it represented during that year.

Frame Hospital Weights. Hospital-level sampling weights were also calculated to represent the entire collection of states in the frame using the same post-stratification scheme as described above for the weights to represent the universe. For each year, within stratum s, each NIS sample hospital's frame weight was calculated as:

$$W_s(\text{frame}) = N_s(\text{frame}) \div N_s(\text{sample}).$$

N_s(frame) was the total number of universe community hospitals within stratum s in the states that contributed data to the frame. N_s(sample) was the number of sample hospitals selected for the NIS in stratum s. Thus, each hospital's frame weight is equal to the number of universe hospitals it represented in the frame states during that year.

State Hospital Weights. For each year, a hospital's weight to its state was calculated in a similar fashion. Within each state, strata often had to be collapsed after sample selection for development of weights to ensure a minimum of two sample hospitals within each stratum. For each state and each year, within stratum s, each NIS sample hospital's state weight was calculated as:

$$W_s(\text{state}) = N_s(\text{state}) \div N_s(\text{state sample}).$$

N_s(state) was the number of universe community hospitals in the state within stratum s. N_s(state sample) was the number of hospitals selected for the NIS from that state in stratum s. Thus, each hospital's state weight is equal to the number of hospitals that it represented in its state during that year.

All of these hospital weights can be rescaled if necessary for selected analyses, to sum to the NIS hospital sample size each year.

Discharge-Level Sampling Weights

The calculations for discharge-level sampling weights were very similar to the calculations of hospital-level sampling weights. The discharge weights usually are constant for all discharges within a stratum.

The only exceptions were for strata with sample hospitals that, according to the AHA files, were open for the entire year but contributed less than their full year of data to the NIS. For those hospitals, we *adjusted* the number of observed discharges by a factor 4 ÷ Q, where Q was the number of calendar quarters that the hospital contributed discharges to the NIS. For example, when a sample hospital contributed only two quarters of discharge data to the NIS, the *adjusted* number of discharges was double the observed number.

With that minor adjustment, each discharge weight is essentially equal to the number of reference (universe, frame, or state) discharges that each sampled discharge represented in its stratum. This calculation was possible because the number of total discharges was available for every hospital in the universe from the AHA files. Each universe hospital's AHA discharge total was calculated as the sum of newborns and total facility discharges.

Universe Discharge Weights. Discharge weights to the universe were calculated by post-stratification. Hospitals were stratified just as they were for universe hospital weight calculations. Within stratum s, for hospital i, each NIS sample discharge's universe weight was calculated as:

$$DW_{is}(\text{universe}) = [DN_s(\text{universe}) \div ADN_s(\text{sample})] * (4 \div Q_i),$$

where DN_s(universe) was the number of discharges from community hospitals in the universe within stratum s; ADN_s(sample) was the number of *adjusted* discharges from sample hospitals selected for the NIS; and Q_i was the number of quarters of discharge data contributed by hospital i to the NIS (usually $Q_i = 4$). Thus, each discharge's weight is equal to the number of universe discharges it represented in stratum s during that year.

Frame Discharge Weights. Discharge-level sampling weights were also calculated to represent all discharges from the entire collection of states in the frame using the same post-stratification scheme described above for the discharge weights to represent the universe. For each year, within stratum s, for hospital i, each NIS sample discharge's frame weight was calculated as:

$$W_{is}(\text{frame}) = [DN_s(\text{frame}) \div ADN_s(\text{sample})] * (4 \div Q_i),$$

DN_s(frame) was the number of discharges from all community hospitals in the states that contributed to the frame within stratum s. ADN_s(sample) was the number of *adjusted* discharges from sample hospitals selected for the NIS in stratum s. Q_i was the number of quarters of discharge data contributed by hospital i to the NIS (usually $Q_i = 4$). Thus, each discharge's frame weight is equal to the number of discharges it represented in the frame states during that year.

State Discharge Weights. A discharge's weight to its state was similarly calculated. Strata were collapsed in the same way as they were for the state hospital weights to ensure a minimum of two sample hospitals within each stratum. Within stratum s, for hospital i, each NIS sample discharge's state weight was calculated as:

$$W_{is}(\text{state}) = [DN_s(\text{state}) \div ADN_s(\text{state sample})] * (4 \div Q_i),$$

DN_s(state) was the number of discharges from all community hospitals in the state within stratum s. ADN_s(state sample) was the *adjusted* number of discharges from hospitals selected for the NIS from that state in stratum s. Q_i was the number of quarters of discharge data contributed by hospital i to the NIS (usually $Q_i = 4$). Thus, each discharge's state weight is equal to the number of discharges that it represented in its state during that year.

All of these discharge weights can be rescaled if necessary for selected analyses, to sum to the NIS discharge sample size each year.

Discharge Weights for 10 Percent Subsamples

In the 10 percent subsamples, each discharge had a 10 percent chance of being drawn. Therefore, the discharge weights contained in the Hospital Weights file can be multiplied by 10 for each of the subsamples, or multiplied by 5 for the two subsamples combined.

DATA ANALYSIS

Variance Calculations

It may be important for researchers to calculate a measure of precision for some estimates based on the NIS sample data. Variance estimates must take into account both the sampling design and the form of the statistic. The sampling design was a stratified, single-stage cluster sample. A stratified random sample of hospitals (clusters) was drawn and then *all* discharges were included from each selected hospital.

If hospitals inside the frame were similar to hospitals outside the frame, the sample hospitals can be treated as if they were randomly selected from the entire universe of hospitals within each stratum. Standard formulas for a stratified, single-stage cluster sampling without replacement could be used to calculate statistics and their variances in most applications.

A multitude of statistics can be estimated from the NIS data. Several computer programs are listed below that calculate statistics and their variances from sample survey data. Some of these programs use general methods of variance calculations (e.g., the jackknife and balanced half-sample replications) that take into account the sampling design. However, it may be desirable to calculate variances using formulas specifically developed for some statistics.

In most cases, computer programs are readily available to perform these calculations. For instance, OSIRIS IV, developed at the University of Michigan, and SUDAAN, developed at the Research Triangle Institute, do calculations for numerous statistics arising from the stratified, single-stage cluster sampling design. An example of how to use SUDAAN to calculate variances in the NIS is presented in Technical Supplement: *Calculating Variances Using Data from the HCUP Nationwide Inpatient Sample.*

These variance calculations are based on finite-sample theory, which is an appropriate method for obtaining cross-sectional, nationwide estimates of outcomes. According to finite-sample theory, the intent of the estimation process is to obtain estimates that are precise representations of the nationwide population at a specific point in time. In the context of the NIS, any estimates that attempt to accurately describe characteristics (such as expenditure and utilization patterns or hospital market factors) and interrelationships among characteristics of hospitals and discharges during a specific year from 1988 to 1993 should be governed by finite-sample theory.

Alternatively, in the study of hypothetical population outcomes not limited to a specific point in time, analysts may be less interested in specific characteristics from the finite population (and time period) from which the *sample* was drawn, than they are in hypothetical characteristics of a conceptual "superpopulation" from which any particular finite *population* in a given year might have been drawn. According to this superpopulation model, the nationwide population in a given year is only a snapshot in time of the possible interrelationships among hospital, market, and discharge characteristics. In a given year, all possible interactions between such characteristics may not have been observed, but analysts may wish to predict or simulate interrelationships that may occur in the future.

Under the finite-population model, the variances of estimates approach zero as the sampling fraction approaches one, since the population is defined at that point in time, and because the estimate is for a characteristic as it existed at the time of sampling. This is in contrast to the superpopulation model, which adopts a stochastic viewpoint rather than a deterministic viewpoint. That is, the nationwide population in a particular year is viewed as a random sample of some underlying superpopulation over time.

Different methods are used for calculating variances under the two sample theories. Under the superpopulation (stochastic) model, procedures (such as those described by Potthoff, Woodbury, and Manton[2]) have been developed to draw inferences using weights from complex samples. In this context, the survey weights are not used to weight the sampled cases to the universe, because the universe is conceptually infinite in size. Instead, these weights are used to produce unbiased estimates of parameters that govern the superpopulation.

In summary, the choice of an appropriate method for calculating variances for nationwide estimates depends on the type of measure and the intent of the estimation process.

Computer Software for Variance Calculations

The hospital weights will be useful for producing hospital-level statistics for analyses that use the *hospital* as the unit of analysis, and the discharge weights will be useful for producing discharge-level statistics for analyses that use the *discharge* as the unit of analysis. These would be used to weight the sample data in estimating population statistics.

Several statistical programming packages allow weighted analyses.[3] For example, nearly all SAS (Statistical Analysis System) procedures can incorporate weights.

In addition, several publicly available subroutines have been developed specifically for calculating statistics and their standard errors from survey data:

- OSIRIS IV was developed by L. Kish, N. Van Eck, and M. Frankel at the Survey Research Center, University of Michigan. It consists of two main programs for estimating variances from complex survey designs.

- SUDAAN, a set of SAS subroutines, was developed at the Research Triangle Institute by B. V. Shah. It is adequate for handling most survey designs with stratification. The procedures can handle estimation and variance estimation for means, proportions, ratios, and regression coefficients.

- SUPER CARP (Cluster Analysis and Regression Program) was developed at Iowa State University by W. Fuller, M. Hidiroglou, and R. Hickman. This program computes estimates and variance estimates for multistage, stratified sampling designs with arbitrary probabilities of selection. It can handle estimated totals, means, ratios, and regression estimates.

The NIS database includes a Hospital Weights file with variables required by these programs to calculate finite population statistics. In addition to the sample weights described earlier, hospital identifiers (PSUs), stratification variables, and stratum-specific totals for the numbers of discharges and hospitals are included so that finite-population corrections (FPCs) can be applied to variance estimates.

In addition to these subroutines, standard errors can be estimated by validation and cross-validation techniques. Given that a very large number of observations will be available for most analyses, it may be feasible to set aside a part of the data for validation purposes. Standard errors and confidence intervals can then be calculated from the validation data. If the analytical file is too small to set aside a large validation sample, cross-validation techniques may be used.

For example, tenfold cross-validation would split the data into ten equal-sized subsets. The estimation would take place in ten iterations. At each iteration, the outcome of interest is predicted for one-tenth of the observations by an estimate based on a model fit to the other nine-tenths of the observations. Unbiased estimates of error variance are then obtained by comparing the actual values to the predicted values obtained in this manner.

Finally, it should be noted that a large array of hospital-level variables are available for the entire universe of hospitals, including those outside the sampling frame. For instance, the variables from the AHA surveys and from the Medicare Cost Reports are available for nearly all hospitals. To the extent that hospital-level outcomes correlate with these variables, they may be used to sharpen regional and nationwide estimates.

As a simple example, each hospital's number of C-sections would be correlated with their total number of deliveries. The number of C-sections must be obtained from discharge data, but the number of deliveries is available from AHA data. Thus, if a regression can be fit predicting C-sections from deliveries based on the NIS data, that regression can then be used to obtain hospital-specific estimates of the number of C-sections for all hospitals in the universe.

Longitudinal Analyses

As previously shown in Table 20, hospitals that continue in the NIS for multiple consecutive years are a subset of the hospitals in the NIS for any one of those years. Consequently, longitudinal analyses of hospital-level outcomes may be biased if they are based on any subset of NIS hospitals limited to continuous NIS membership. In particular, such subsets would tend to contain fewer hospitals that opened, closed, split, merged, or changed strata. Further, the sample weights were developed as annual, cross-sectional weights rather than longitudinal weights. Therefore, different weights might be required, depending on the statistical methods employed by the analyst.

One approach to consider in hospital-level longitudinal analyses is to use repeated-measure models that allow hospitals to have missing values for some years. However, the data are not actually missing for some hospitals, such as those that closed during the study period. In any case, the analyses may be more efficient (e.g., produce more precise estimates) if they account for the potential correlation between repeated measures on the same hospital over time, yet incorporate data from all hospitals in the sample during the study period.

Discharge Subsamples

The two nonoverlapping 10 percent subsamples of discharges were drawn from the NIS file for each year for several reasons pertaining to data analysis. One reason for creating the subsamples was to reduce processing costs for selected studies that will not require the entire NIS. Another reason is that the two subsamples may be used to validate models and obtain unbiased estimates of standard errors. That is, one subsample may be used to estimate statistical models, and the other subsample may be used to test the fit of those models on new data. This is a very important analytical step, particularly in exploratory studies, where one runs the risk of fitting noise.

For example, it is well known that the percentage of variance explained by a regression, R^2, is generally overestimated by the data used to fit a model. The regression model could be estimated from the first subsample and then applied to the second subsample. The squared correlation between the actual and predicted value in the second subsample is an unbiased estimate of the model's true explanatory power when applied to new data.

ENDNOTES

1. Coffey, R. and D. Farley (1988, July). *HCUP-2 Project Overview,* (DHHS Publication No. (PHS) 88-3428. Hospital Studies Program Research Note 10, National Center for Health Services Research and Health Care Technology Assessment, Rockville, MD: Public Health Service.

2. Potthoff, R.F., M.A. Woodbury, and K.G. Manton (1992). "Equivalent Sample Size" and "Equivalent Degrees of Freedom" Refinements for Inference Using Survey Weights Under Superpopulation Models. *Journal of the American Statistical Association*, Vol. 87, 383-396.

3. Carlson, B.L., A.E. Johnson, and S.B. Cohen (1993). An Evaluation of the Use of Personal Computers for Variance Estimation with Complex Survey Data. *Journal of Official Statistics*, Vol. 9, No. 4, 795-814.

TECHNICAL SUPPLEMENT 7:
DESIGN OF THE HCUP NATIONWIDE INPATIENT SAMPLE, RELEASE 3

INTRODUCTION

The Nationwide Inpatient Sample (NIS) of the Healthcare Cost and Utilization Project (HCUP) was established to provide analyses of hospital utilization across the United States. The NIS, Release 1 covers calendar years 1988-1992. The NIS, Release 2 covers calendar year 1993, and the NIS, Release 3 covers calendar year 1994. The target universe includes all acute-care discharges from all community hospitals in the United States; the NIS comprises all discharges from a sample of hospitals in this target universe.

This third release of the NIS contains 6.4 million discharges from a sample of 904 hospitals in 17 states. The first release (1988 through 1992) contains 5.2 to 6.2 million discharges per year from a sample of 758 to 875 hospitals per year in 11 states (8 states for 1988). The second release of the NIS contains 6.5 million discharges from a sample of 913 hospitals in 17 states. Thus, the NIS supports both cross-sectional and longitudinal analyses.

Potential research issues focus on both discharge- and hospital-level outcomes. Discharge outcomes of interest include trends in inpatient treatments with respect to:

- frequency,
- costs,
- lengths of stay,
- effectiveness,
- appropriateness, and
- access to hospital care.

Hospital outcomes of interest include:

- mortality rates,
- complication rates,
- patterns of care,
- diffusion of technology, and
- trends toward specialization.

These and other outcomes are of interest for the nation as a whole and for policy-relevant inpatient subgroups defined by geographic regions, patient demographics, hospital characteristics, physician characteristics, and pay sources.

This report provides a detailed description of the NIS, Release 3 sample design, as well as a summary of the resultant hospital sample. Sample weights were developed to obtain national estimates of hospital and inpatient parameters. These weights and other special-use weights are described in detail. Tables include cumulative information for NIS, Release 1 (1988 through 1992); NIS, Release 2 (1993); and NIS, Release 3 (1994) to provide a longitudinal view of the database.

THE NIS HOSPITAL UNIVERSE

The hospital universe is defined by all hospitals that were open during any part of the calendar year and were designated as community hospitals in the American Hospital Association (AHA) Annual Survey of Hospitals. For purposes of the NIS, the definition of a community hospital is that used by the AHA: "all nonfederal short-term general and other specialty hospitals, excluding hospital units of institutions." Consequently, Veterans Hospitals and other federal hospitals are excluded. Table 21 shows the number of universe hospitals for each year based on the AHA Annual Survey.

Table 21. Hospital Universe[1]

Year	Number of Hospitals
1988	5,607
1989	5,548
1990	5,468
1991	5,412
1992	5,334
1993	5,313
1994	5,290

Hospital Merges, Splits, and Closures

All hospital entities that were designated community hospitals in the AHA hospital file were included in the hospital universe. Therefore, if two or more community hospitals merged to create a new community hospital, the original hospitals and the newly-formed hospital were all considered separate hospital entities in the universe for the year of the merge. Likewise, if a community hospital split, the original hospital and all newly created community hospitals were separate entities in the universe for the year of the split. Finally, community hospitals that closed during a year were included as long as they were in operation during some part of the calendar year.

Stratification Variables

To help ensure representativeness, sampling strata were defined based on five hospital characteristics contained in the AHA hospital files. The stratification variables were as follows:

1) *Geographic Region – Northeast, Midwest, West, and South*. This is an important stratifier because practice patterns have been shown to vary substantially by region. For example, lengths of stay tend to be longer in East Coast hospitals than in West Coast hospitals.

2) *Control – government nonfederal, private not-for-profit, and private investor-owned.* These types of hospitals tend to have different missions and different responses to government regulations and policies.

3) *Location – urban or rural.* Government payment policies often differ according to this designation. Also, rural hospitals are generally smaller and offer fewer services than urban hospitals.

4) *Teaching Status – teaching or nonteaching*. The missions of teaching hospitals differ from nonteaching hospitals. In addition, financial considerations differ between these two hospital groups. Currently, the Medicare DRG payments are uniformly higher to teaching hospitals than to nonteaching hospitals. A hospital is considered to be a teaching hospital if it has an AMA-approved residency program or is a member of the Council of Teaching Hospitals (COTH).

5) *Bedsize – small, medium, and large.* Bedsize categories are based on hospital beds, and are specific to the hospital's location and teaching status, as shown in Table 22.

Table 22. Bedsize Categories

Location and Teaching Status	**Hospital Bedsize**		
	Small	**Medium**	**Large**
Rural	1-49	50-99	100+
Urban, nonteaching	1-99	100-199	200+
Urban, teaching	1-299	300-499	500+

Rural hospitals were not split according to teaching status, because rural teaching hospitals were rare. For example, in 1988 there were only 20 rural teaching hospitals. The bedsize categories were defined within location and teaching status because they would otherwise have been redundant. Rural hospitals tend to be small; urban nonteaching hospitals tend to be medium-sized; and urban teaching hospitals tend to be large. Yet it was important to recognize gradations of size within these types of hospitals.

For example, in serving rural discharges, the role of "large" rural hospitals (particularly rural referral centers) often differs from the role of "small" rural hospitals. The cut-off points for the bedsize categories are consistent with those used in *Hospital Statistics,* published annually by the AHA.

To further ensure geographic representativeness, implicit stratification variables included state and three-digit zip code (the first three digits of the hospital's five-digit zip code). The hospitals were sorted according to these variables prior to systematic sampling.

HOSPITAL SAMPLING FRAME

For each year, the *universe* of hospitals was established as all community hospitals located in the U.S. However, it was not feasible to obtain and process all-payer discharge data from a random sample of the entire universe of hospitals for at least two reasons. First, all-payer discharge data were not available from all hospitals for research purposes. Second, based on the experience of prior hospital discharge data collections, it would have been too costly to obtain data from individual hospitals, and it would have been too burdensome to process each hospital's unique data structure.

Therefore, the NIS *sampling frame* was constructed from the subset of universe hospitals that released their discharge data for research use. Two sources for all-payer discharge data were state agencies and private data organizations, primarily state hospital associations. At the time when the sample was drawn, the Agency for Health Care Policy and Research (AHCPR) had agreements with 22 data sources that maintain statewide, all-payer discharge data files to include their data in the HCUP database. However, only 8 states in 1988 and 11 states in 1989-1992 could be included in the first release of the NIS, and an additional 6 states have been included in the second and the third release of the NIS, as shown in Table 23.

Table 23. States in the Frame for the NIS, Release 1, NIS, Release 2, and NIS, Release 3

Years	States in the Frame
NIS, Release 1	
1988	California, Colorado, Florida, Iowa, Illinois, Massachusetts, New Jersey, and Washington
1989-1992	Add Arizona, Pennsylvania, and Wisconsin
NIS, Release 2	
1993	Add Connecticut, Kansas, Maryland, New York, Oregon, South Carolina
NIS, Release 3	
1994	Add Connecticut, Kansas, Maryland, New York, Oregon, South Carolina

The list of the entire frame of hospitals was composed of all AHA community hospitals in each of the frame states *that could be matched to the discharge data provided to HCUP*, with restrictions on the hospitals that could be included from Illinois and South Carolina. If an AHA community hospital could not be matched to the discharge data provided by the data source, it was eliminated from the sampling frame (but not from the universe).

The Illinois Health Care Cost Containment Council stipulated that no more than 40 percent of the data provided by Illinois could be included in the database for any calendar quarter. As a result, the number of Illinois community hospitals in the frame was restricted, and 104 of the 208 Illinois community hospitals in the universe (50 percent of hospitals) were randomly selected using the same methodology used to select the NIS hospital sample. That is, Illinois hospitals were stratified on the stratification variables described above, and a systematic random sample of hospitals was drawn for the frame. This prevented the sample from including more than 40 percent of Illinois discharges.

South Carolina stipulated that only hospitals that appear in sampling strata with two or more hospitals were to be included in the NIS. Four South Carolina hospitals were excluded from the frame since there were fewer than two South Carolina hospitals in four sampling frame strata. The remaining 59 South Carolina community hospitals are included in the frame.

The number of frame hospitals for each year is shown in Table 24.

Table 24. Hospital Frame

Year	Number of Hospitals
1988	1,247
1989	1,658
1990	1,620
1991	1,604
1992	1,591
1993	2,168
1994	2,135

HOSPITAL SAMPLE DESIGN

Design Requirements

The NIS is a stratified probability sample of hospitals in the frame, with sampling probabilities calculated to select 20 percent of the universe contained in each stratum. The overall objective was to select a sample of hospitals "generalizable" to the target universe, which includes hospitals outside the frame (zero probability of selection). Moreover, this sample was to be geographically dispersed, yet drawn from the subset of states with inpatient discharge data that agreed to provide such data to the project.

It should be possible, for example, to estimate DRG-specific average lengths of stay over all U.S. hospitals using weighted average lengths of stay, based on averages or regression estimates from the NIS. Ideally, relationships among outcomes and their correlates estimated from the NIS should generally hold across all U.S. hospitals. However, since only 17 states contributed data to this third release, some estimates may differ from estimates from comparative data sources. When possible, estimates based on the NIS should be checked against national benchmarks, such as Medicare data or data from the National Hospital Discharge Survey to determine the appropriateness of the NIS for specific analyses.

The target sample size was 20 percent of the total number of community hospitals in the U.S. for 1994. This sample size was determined by AHCPR based on their experience with similar research databases.

Alternative stratified sampling allocation schemes were considered. However, allocation proportional to the number of hospitals is preferred for several reasons:

- Fewer than 10 percent of government-planned database applications will produce nationwide estimates. The major government applications will investigate relationships among variables. For example, government researchers will do a substantial amount of regression modeling with these data.

- The HCUP-2 sample[2] used the same stratification and allocation scheme, and it has served AHCPR analysts well. Moreover, the large number of sample hospitals and discharges seemingly reduced the need for variance-reducing allocation schemes.

- AHCPR researchers wanted a simple, easily understood sampling methodology. It was an appealing idea that the NIS sample could be a "miniaturization" of the universe of hospitals (with the obvious geographical limitations imposed by data availability).

- AHCPR statisticians considered other optimal allocation schemes, including sampling hospitals with probabilities proportional to size (number of discharges), and they concluded that sampling with probability proportional to the number of hospitals was preferable. Even though it was recognized that the approach chosen would not be as efficient, the extremely large sample sizes yield good estimates. Furthermore, because the data are to be used for purposes other than producing national estimates, it is critical that all hospital types (including small hospitals) are adequately represented.

Hospital Sampling Procedure

Once the universe of hospitals was stratified, up to 20 percent of the total number of U.S. hospitals was randomly selected within each stratum. If too few frame hospitals were in the stratum, then all frame hospitals were selected for the NIS, subject to sampling restrictions specified by states. To simplify variance calculations, at least two hospitals were drawn from each stratum. If fewer than two frame hospitals were contained in a stratum, then that stratum was merged with an "adjacent" stratum containing hospitals with similar characteristics.

A systematic random sample was drawn from each stratum, after sorting hospitals by state within each stratum, then by the three-digit zip code (the first three digits of the hospital's five-digit zip code) within each state, and then by a random number within each three-digit zip code. These sorts ensured further geographic generalizability of hospitals within the frame states, and random ordering of hospitals within three-digit zip codes.

Generally, three-digit zip codes that are near in value are geographically near within a state. Furthermore, the U.S. Postal Service locates regional mail distribution centers at the three-digit level. Thus, the boundaries tend to be a compromise between geographic size and population size.

1994 NIS Hospital Sampling Procedure

The 1994 sample was drawn by a procedure that retained most of the 1993 hospitals, while allowing hospitals new to the frame an opportunity to enter the 1994 NIS.

Even in frame states that were present in the 1993 sample, hospitals that opened in 1994 needed a chance to enter the sample. Also, hospitals that changed strata between 1993 and 1994 were considered new to the 1994 frame.

Consequently, a recursive procedure was developed to update the sample from year to year in a way that properly accounted for changes in stratum size, composition, and sampling rate. The goal of this procedure was to maximize the year-to-year overlap among sample hospitals, yet keep the sampling rate constant for all hospitals *within a stratum*.

The following procedure provides rules for creating a "year 2" sample, given that a "year 1" sample had already been drawn. In this example, year 1 would be 1993 and year 2 would be 1994. All notation is assumed to refer to sizes and probabilities within a particular stratum.

Probabilities P_1 and P_2 were calculated for sampling hospitals from the frame within the stratum for year 1 and year 2, respectively, based on the frame and universe for year 1 and year 2, respectively. These probabilities were set by the same algorithm used to calculate P for the 1988 hospital sample (see Technical Supplement: *Design of the HCUP Nationwide Inpatient Sample, Release 1*, section "1988 NIS Hospital Sampling Procedure.")

Now consider the three possibilities associated with changes between years 1 and 2 in the stratum-specific hospital sampling probabilities:

1. $P_2 = P_1$: The target probability was unchanged.

2. $P_2 < P_1$: The target probability decreased.

3. $P_2 > P_1$: The target probability increased.

Below is the procedure used for each of these three cases with one exception: if the stratum-specific probability of selection P_2 was equal to 1, then all frame hospitals were selected for the year 2 sample, regardless of the value of P_1.

Stratum-Specific Sampling Rates the Same ($P_2 = P_1$). If the probability P_2 was the same as P_1, all hospitals in the year 1 sample that remained in the year 2 frame were retained for the year 2 sample. Any new frame hospitals (those in the year 2 frame but not in the year 1 frame) were selected at the rate P_2, using the systematic sampling method described for the 1988 sample selection in Technical Supplement: *Design of the HCUP Nationwide Inpatient Sample, Release 1*.

Stratum-Specific Sampling Rate Decreased ($P_2 < P_1$). Now consider the case where the probability of selection decreased between years 1 and 2. First, hospitals new to the frame were sampled with probability P_2. Second, hospitals previously selected for the year 1 sample (that remained in the year 2 frame) were selected for the year 2 sample with probability $P_2 \div P_1$.

The justification for this second procedure was straightforward. For the year 1 sample hospitals that stayed in the frame, the year 1 sample was viewed as the first stage of a two-stage sampling process. The first stage was carried out at the sampling rate of P_1. The second stage was carried out at the sampling rate of $P_2 \div P_1$. Consequently, the "overall" probability of selection was $P_1 \times P_2 \div P_1 = P_2$.

Stratum-Specific Sampling Rate Increased ($P_2 > P_1$). The procedures associated with the case in which the probability of selection was increased between year 1 and year 2 were equally straightforward. First, hospitals new to the frame were sampled with probability P_2. Second, hospitals that were selected in year 1 (that remained in the year 2 frame) were selected for the year 2 sample. Third, hospitals that were in the frame for both years 1 and 2, but not selected for the year 1 sample, were selected for the year 2 sample with probability $(P_2\text{-}P_1) \div (1\text{-}P_1)$.

The justification for this sampling rate, $(P_2\text{-}P_1) \div (1\text{-}P_1)$, is somewhat complex. In year 1 certain frame hospitals were included in the sample at the rate P_1. This can also be viewed as having excluded a set of hospitals at the rate $(1\text{-}P_1)$. Likewise, in year 2 it was imperative that each hospital excluded from the year 1 sample be excluded from the year 2 sample at an overall rate of $(1-P_2)$.

Since $P_2 > P_1$, then $(1-P_2) < (1-P_1)$. Therefore, just as was done for the case of $P_2 < P_1$, multistage selection was implemented. However, it was implemented for exclusion rather than inclusion.

Therefore, those hospitals excluded from the year 1 sample were also excluded from the year 2 sample at the rate $S = (1-P_2) \div (1-P_1)$. This gave them the desired overall *exclusion* rate of $(1-P_1) \times (1-P_2) \div (1-P_1) = (1-P_2)$. Consequently, the *inclusion* rate for these hospitals was set at $1-S = (P_2-P_1) \div (1-P_1)$.

Zero-Weight Hospitals

The 1994 sample contains no zero-weight hospitals. For a description of zero-weight hospitals in the 1988-1992 sample, see the Technical Supplement: *Design of the HCUP Nationwide Inpatient Sample, Release 1*.

Ten Percent Subsamples

Two nonoverlapping 10 percent subsamples of discharges were drawn from the NIS file for each year. The subsamples were selected by drawing every tenth discharge starting with two different starting points (randomly selected between 1 and 10). Having a different starting point for each of the two subsamples guaranteed that they would not overlap. Discharges were sampled so that 10 percent of each hospital's discharges in each quarter were selected for each of the subsamples. The two samples can be combined to form a single, generalizable 20 percent subsample of discharges.

FINAL HOSPITAL SAMPLE

The annual numbers of hospitals and discharges in NIS, Release 1; NIS, Release 2; and NIS Release 3 are shown in Table 25, for both the regular NIS sample and the total sample (which includes zero-weight hospitals for 1988-1992).

Table 25. NIS Hospital Sample

Year	Regular Sample		Total Sample	
	Number of Hospitals	Number of Discharges	Number of Hospitals	Number of Discharges
NIS, Release 1				
1988	758	5,242,904	759	5,265,756
1989	875	6,067,667	882	6,110,064
1990	861	6,156,638	871	6,268,515
1991	847	5,984,270	859	6,156,188
1992	838	6,008,001	856	6,195,744
NIS, Release 2				
1993	913	6,538,976	913	6,538,976
NIS, Release 3				
1994	904	6,385,011	904	6,385,011
Total		42,383,467		42,920,254

A more detailed breakdown of the 1994 NIS hospital sample by geographic region is shown in Table 26. For each geographic region, Table 26 shows the number of:

- universe hospitals (Universe),
- frame hospitals (Frame),
- sampled hospitals (Sample),
- target hospitals (Target = 20 percent of the universe), and
- shortfall hospitals (Shortfall = Sample - Target).

Table 26. Number of Hospitals in Universe, Frame, Regular Sample, Target, and Shortfall By Region, 1994

Region	Universe	Frame	Sample	Target	Shortfall
NE	780	654	168	156	12
MW	1,527	473	304	305	-1
S	2,010	313	256	403	-147
W	973	695	176	195	-19
Total	5,290	2,135	904	1,059	-155

For example, in 1994 the Northeast region contained 780 hospitals in the universe. It also contained 654 hospitals in the frame, of which 168 hospitals were drawn for the sample. This was 12 hospitals more than the target sample size of 156.

Table 27 shows the number of hospitals in the universe, frame, and regular sample for each state in the sampling frame for 1994. In all states except Illinois and South Carolina, the difference between the universe and the frame represents the difference in the number of community hospitals in the 1994 AHA Annual Survey of Hospitals and the number of community hospitals for which data were supplied to HCUP. As explained earlier, the number of hospitals in the Illinois frame is approximately 50 percent of the hospitals in the Illinois universe in order to comply with the agreement with the data source concerning the restriction on the number of Illinois discharges. The number of hospitals in the South Carolina frame is eight fewer than the South Carolina universe. Four hospitals were excluded because of sampling restrictions stipulated by South Carolina, and four hospitals were not included in the data supplied to HCUP.

The number of hospitals in the NIS hospital samples that continue across multiple sample years is shown in Table 28. This table will be of interest to those who may combine Release 1, 2, and 3 of the NIS. Table 28 shows that longitudinal cohorts that span several years and include 1988 and 1993 are the lowest in number of continuing sample hospitals. For example, if 1988 is taken as a starting year, only 44.2 percent of the 1988 hospital sample continued in the 1994 sample (335 of 758).

Table 27. Number of Hospitals in the Universe, Frame, and Regular Sample for States in the Sampling Frame: 1994

State	Universe	Frame	Sample
AZ	61	49	12
CA	430	428	102
CO	69	68	22
CT	37	32	7
FL	220	204	163
IA	116	116	64
IL	208	104	77
KS	137	126	71
MA	95	84	27
MD	50	50	42
NJ	94	86	19
NY	230	228	62
OR	63	62	19
PA	230	224	53
SC	67	59	51
WA	90	88	21
WI	127	127	92
Total	2324	2135	904

Table 28. Number of Hospitals and Discharges in Longitudinal Cohort

Number of Years	Calendar Years	Longitudinal Regular Sample Hospitals	% of Base Year Sample	Longitudinal Regular Sample Discharges
2	1988-1989	610	80.5	8,492,039
	1989-1990	815	93.1	11,525,749
	1990-1991	802	93.1	11,297,175
	1991-1992	781	92.2	11,272,981
	1992-1993	609	72.7	8,804,638
	1993-1994	693	75.9	10,271,404
3	1988-1990	573	75.6	12,168,677
	1989-1991	763	87.2	16,074,381
	1990-1992	745	86.5	16,085,651
	1991-1993	570	67.3	12,559,421
	1992-1994	540	64.4	11,279,667
4	1988-1991	542	71.5	15,096,807
	1989-1992	709	81.0	20,340,970
	1990-1993	548	63.6	16,023,500
	1991-1994	508	60.0	14,481,319
5	1988-1992	502	66.2	18,106,098
	1989-1993	523	59.8	19,000,777
	1990-1994	490	56.9	17,437,229
6	1988-1993	378	49.9	16,906,818
	1989-1994	471	53.8	19,987,910
7	1988-1994	335	44.2	17,128,064

SAMPLING WEIGHTS

Although the sampling design was simple and straightforward, it is necessary to incorporate sample weights to obtain state and national estimates. Therefore, sample weights were developed separately for hospital- and discharge-level analyses. Three hospital-level weights were developed to weight NIS sample hospitals to the state, frame, and universe. Similarly, three discharge-level weights were developed to weight NIS sample discharges to the state, frame, and universe.

Hospital-Level Sampling Weights

Universe Hospital Weights. Hospital weights to the universe were calculated by post-stratification. For each year, hospitals were stratified on the same variables that were used for sampling: geographic region, urban/rural location, teaching status, bedsize, and control. The strata that were collapsed for sampling were also collapsed for sample weight calculations. Within stratum s, each NIS sample hospital's universe weight was calculated as:

$$W_s(\text{universe}) = N_s(\text{universe}) \div N_s(\text{sample}),$$

where N_s(universe) and N_s(sample) were the number of community hospitals within stratum s in the universe and sample, respectively. Thus, each hospital's universe weight is equal to the number of universe hospitals it represented during that year.

Frame Hospital Weights. Hospital-level sampling weights were also calculated to represent the entire collection of states in the frame using the same post-stratification scheme as described above for the weights to represent the universe. For each year, within stratum s, each NIS sample hospital's frame weight was calculated as:

$$W_s(\text{frame}) = N_s(\text{frame}) \div N_s(\text{sample}).$$

N_s(frame) was the total number of universe community hospitals within stratum s in the states that contributed data to the frame. N_s(sample) was the number of sample hospitals selected for the NIS in stratum s. Thus, each hospital's frame weight is equal to the number of universe hospitals it represented in the frame states during that year.

State Hospital Weights. For each year, a hospital's weight to its state was calculated in a similar fashion. Within each state, strata often had to be collapsed after sample selection for development of weights to ensure a minimum of two sample hospitals within each stratum. For each state and each year, within stratum s, each NIS sample hospital's state weight was calculated as:

$$W_s(\text{state}) = N_s(\text{state}) \div N_s(\text{state sample}).$$

N_s(state) was the number of universe community hospitals in the state within stratum s. N_s(state sample) was the number of hospitals selected for the NIS from that state in stratum s. Thus, each hospital's state weight is equal to the number of hospitals that it represented in its state during that year.

All of these hospital weights can be rescaled if necessary for selected analyses, to sum to the NIS hospital sample size each year.

Discharge-Level Sampling Weights

The calculations for discharge-level sampling weights were very similar to the calculations of hospital-level sampling weights. The discharge weights usually are constant for all discharges within a stratum.

The only exceptions were for strata with sample hospitals that, according to the AHA files, were open for the entire year but contributed less than their full year of data to the NIS. For those hospitals, we *adjusted* the number of observed discharges by a factor 4 ÷ Q, where Q was the number of calendar quarters that the hospital contributed discharges to the NIS. For example, when a sample hospital contributed only two quarters of discharge data to the NIS, the *adjusted* number of discharges was double the observed number.

With that minor adjustment, each discharge weight is essentially equal to the number of reference (universe, frame, or state) discharges that each sampled discharge represented in its stratum. This calculation was possible because the number of total discharges was available for every hospital in the universe from the AHA files. Each universe hospital's AHA discharge total was calculated as the sum of newborns and total facility discharges.

Universe Discharge Weights. Discharge weights to the universe were calculated by post-stratification. Hospitals were stratified just as they were for universe hospital weight calculations. Within stratum s, for hospital i, each NIS sample discharge's universe weight was calculated as:

$$DW_{is}(\text{universe}) = [DN_s(\text{universe}) \div ADN_s(\text{sample})] * (4 \div Q_i),$$

where DN_s(universe) was the number of discharges from community hospitals in the universe within stratum s; ADN_s(sample) was the number of *adjusted* discharges from sample hospitals selected for the NIS; and Q_i was the number of quarters of discharge data contributed by hospital i to the NIS (usually $Q_i = 4$). Thus, each discharge's weight is equal to the number of universe discharges it represented in stratum s during that year.

Frame Discharge Weights. Discharge-level sampling weights were also calculated to represent all discharges from the entire collection of states in the frame using the same post-stratification scheme described above for the discharge weights to represent the universe. For each year, within stratum s, for hospital i, each NIS sample discharge's frame weight was calculated as:

$$W_{is}(\text{frame}) = [DN_s(\text{frame}) \div ADN_s(\text{sample})] * (4 \div Q_i),$$

DN_s(frame) was the number of discharges from all community hospitals in the states that contributed to the frame within stratum s. ADN_s(sample) was the number of *adjusted* discharges from sample hospitals selected for the NIS in stratum s. Q_i was the number of quarters of discharge data contributed by hospital i to the NIS (usually $Q_i = 4$). Thus, each discharge's frame weight is equal to the number of discharges it represented in the frame states during that year.

State Discharge Weights. A discharge's weight to its state was similarly calculated. Strata were collapsed in the same way as they were for the state hospital weights to ensure a minimum of two sample hospitals within each stratum. Within stratum s, for hospital i, each NIS sample discharge's state weight was calculated as:

$$W_{is}(\text{state}) = [DN_s(\text{state}) \div ADN_s(\text{state sample})] * (4 \div Q_i),$$

DN_s(state) was the number of discharges from all community hospitals in the state within stratum s. ADN_s(state sample) was the *adjusted* number of discharges from hospitals selected for the NIS from that state in stratum s. Q_i was the number of quarters of discharge data contributed by hospital i to the NIS (usually $Q_i = 4$). Thus, each discharge's state weight is equal to the number of discharges that it represented in its state during that year.

All of these discharge weights can be rescaled if necessary for selected analyses, to sum to the NIS discharge sample size each year.

Discharge Weights for 10 Percent Subsamples

In the 10 percent subsamples, each discharge had a 10 percent chance of being drawn. Therefore, the discharge weights contained in the Hospital Weights file can be multiplied by 10 for each of the subsamples, or multiplied by 5 for the two subsamples combined.

DATA ANALYSIS

Variance Calculations

It may be important for researchers to calculate a measure of precision for some estimates based on the NIS sample data. Variance estimates must take into account both the sampling design and the form of the statistic. The sampling design was a stratified, single-stage cluster sample. A stratified random sample of hospitals (clusters) were drawn and then *all* discharges were included from each selected hospital.

If hospitals inside the frame were similar to hospitals outside the frame, the sample hospitals can be treated as if they were randomly selected from the entire universe of hospitals within each stratum. Standard formulas for a stratified, single-stage cluster sampling without replacement could be used to calculate statistics and their variances in most applications.

A multitude of statistics can be estimated from the NIS data. Several computer programs are listed below that calculate statistics and their variances from sample survey data. Some of these programs use general methods of variance calculations (e.g., the jackknife and balanced half-sample replications) that take into account the sampling design. However, it may be desirable to calculate variances using formulas specifically developed for some statistics.

In most cases, computer programs are readily available to perform these calculations. For instance, OSIRIS IV, developed at the University of Michigan, and SUDAAN, developed at the Research Triangle Institute, do calculations for numerous statistics arising from the stratified, single-stage cluster sampling design. An example of using SUDAAN to calculate variances in the NIS is presented in Technical Supplement: *Calculating Variances Using Data from the HCUP Nationwide Inpatient Sample*.[3]

These variance calculations are based on finite-sample theory, which is an appropriate method for obtaining cross-sectional, nationwide estimates of outcomes. According to finite-sample theory, the intent of the estimation process is to obtain estimates that are precise representations of the nationwide population at a specific point in time. In the context of the NIS, any estimates that attempt to accurately describe characteristics (such as expenditure and utilization patterns or hospital market factors) and interrelationships among characteristics of hospitals and discharges during a specific year from 1988 to 1994 should be governed by finite-sample theory.

Alternatively, in the study of hypothetical population outcomes not limited to a specific point in time, analysts may be less interested in specific characteristics from the finite population (and time period) from which the *sample* was drawn, than they are in hypothetical characteristics of a conceptual "superpopulation" from which any particular finite *population* in a given year might have been drawn. According to this superpopulation model, the nationwide population in a given year is only a snapshot in time of the possible interrelationships among hospital, market, and discharge characteristics. In a given year, all possible interactions between such characteristics may not have been observed, but analysts may wish to predict or simulate interrelationships that may occur in the future.

Under the finite-population model, the variances of estimates approach zero as the sampling fraction approaches one, since the population is defined at that point in time, and because the estimate is for a characteristic as it existed at the time of sampling. This is in contrast to the superpopulation model, which adopts a stochastic viewpoint rather than a deterministic viewpoint. That is, the nationwide population in a particular year is viewed as a random sample of some underlying superpopulation over time.

Different methods are used for calculating variances under the two sample theories. Under the superpopulation (stochastic) model, procedures (such as those described by Potthoff, Woodbury, and Manton[4]) have been developed to draw inferences using weights from complex samples. In this context, the survey weights are not used to weight the sampled cases to the universe, because the universe is conceptually infinite in size. Instead, these weights are used to produce unbiased estimates of parameters that govern the superpopulation.

In summary, the choice of an appropriate method for calculating variances for nationwide estimates depends on the type of measure and the intent of the estimation process.

Computer Software for Variance Calculations

The hospital weights will be useful for producing hospital-level statistics for analyses that use the *hospital* as the unit of analysis, and the discharge weights will be useful for producing discharge-level statistics for analyses that use the *discharge* as the unit of analysis. These would be used to weight the sample data in estimating population statistics.

Several statistical programming packages allow weighted analyses.[5] For example, nearly all SAS (Statistical Analysis System) procedures incorporate weights.

In addition, several publicly available subroutines have been developed specifically for calculating statistics and their standard errors from survey data:

- OSIRIS IV was developed by L. Kish, N. Van Eck, and M. Frankel at the Survey Research Center, University of Michigan. It consists of two main programs for estimating variances from complex survey designs.

- SUDAAN, a set of SAS subroutines, was developed at the Research Triangle Institute by B. V. Shah. It is adequate for handling most survey designs with stratification. The procedures can handle estimation and variance estimation for means, proportions, ratios, and regression coefficients.

- SUPER CARP (Cluster Analysis and Regression Program) was developed at Iowa State University by W. Fuller, M. Hidiroglou, and R. Hickman. This program computes estimates and variance estimates for multistage, stratified sampling designs with arbitrary probabilities of selection. It can handle estimated totals, means, ratios, and regression estimates.

The NIS database includes a Hospital Weights file with variables required by these programs to calculate finite population statistics. In addition to the sample weights described earlier, hospital identifiers (PSUs), stratification variables, and stratum-specific totals for the numbers of discharges and hospitals are included so that finite-population corrections (FPCs) can be applied to variance estimates.

In addition to these subroutines, standard errors can be estimated by validation and cross-validation techniques. Given that a very large number of observations will be available for most analyses, it may be feasible to set aside a part of the data for validation purposes. Standard errors and confidence intervals can then be calculated from the validation data. If the analytical file is too small to set aside a large validation sample, cross-validation techniques may be used.

For example, tenfold cross-validation would split the data into ten equal-sized subsets. The estimation would take place in ten iterations. At each iteration, the outcome of interest is predicted for one-tenth of the observations by an estimate based on a model fit to the other nine-tenths of the observations. Unbiased estimates of error variance are then obtained by comparing the actual values to the predicted values obtained in this manner.

Finally, it should be noted that a large array of hospital-level variables are available for the entire universe of hospitals, including those outside the sampling frame. For instance, the variables from the AHA surveys and from the Medicare Cost Reports are available for nearly all hospitals. To the extent that hospital-level outcomes correlate with these variables, they may be used to sharpen regional and nationwide estimates.

As a simple example, each hospital's number of C-sections would be correlated with their total number of deliveries. The number of C-sections must be obtained from discharge data, but the number of deliveries is available from AHA data. Thus, if a regression can be fit predicting C-sections from deliveries based on the NIS data, that regression can then be used to obtain hospital-specific estimates of the number of C-sections for all hospitals in the universe.

Longitudinal Analyses

As previously shown in Table 28, hospitals that continue in the NIS for multiple consecutive years are a subset of the hospitals in the NIS for any one of those years. Consequently, longitudinal analyses of hospital-level outcomes may be biased if they are based on any subset of NIS hospitals limited to continuous NIS membership. In particular, such subsets would tend to contain fewer hospitals that opened, closed, split, merged, or changed strata. Further, the sample weights were developed as annual, cross-sectional weights rather than longitudinal weights. Therefore, different weights might be required, depending on the statistical methods employed by the analyst.

One approach to consider in hospital-level longitudinal analyses is to use repeated-measure models that allow hospitals to have missing values for some years. However, the data are not actually missing for some hospitals, such as those that closed during the study period. In any case, the analyses may be more efficient (e.g., produce more precise estimates) if they account for the potential correlation between repeated measures on the same hospital over time, yet incorporate data from all hospitals in the sample during the study period.

Discharge Subsamples

The two nonoverlapping 10 percent subsamples of discharges were drawn from the NIS file for each year for several reasons pertaining to data analysis. One reason for creating the subsamples was to reduce processing costs for selected studies that will not require the entire NIS. Another reason is that the two subsamples may be used to validate models and obtain unbiased estimates of standard errors. That is, one subsample may be used to estimate statistical models, and the other subsample may be used to test the fit of those models on new data. This is a very important analytical step, particularly in exploratory studies, where one runs the risk of fitting noise.

For example, it is well known that the percentage of variance explained by a regression, R^2, is generally overestimated by the data used to fit a model. The regression model could be estimated from the first subsample and then applied to the second subsample. The squared correlation between the actual and predicted value in the second subsample is an unbiased estimate of the model's true explanatory power when applied to new data.

ENDNOTES

1. Most AHA surveys do not cover a January-to-December calendar year. The number of hospitals for 1988-1991 are based on the HCUP calendar-year version of the AHA Annual Survey files. To create a calendar-year reporting period, data from the AHA surveys must be apportioned in some manner across calendar years. Survey responses were converted to calendar-year periods for 1988-1991 by merging data from adjacent survey years. The number of hospitals for 1992-1994 are based on the AHA Annual Survey files.

2. Coffey, R. and D. Farley (1988, July). *HCUP-2 Project Overview,* (DHHS Publication No. (PHS) 88-3428. Hospital Studies Program Research Note 10, National Center for Health Services Research and Health Care Technology Assessment, Rockville, MD: Public Health Service.

3. Duffy, S.Q. and J.P. Sommers (1996, March). *Calculating Variances Using Data from the HCUP Nationwide Inpatient Sample*. Rockville, MD: Agency for Health Care Policy and Research.

4. Potthoff, R.F., M.A. Woodbury, and K.G. Manton (1992). "Equivalent Sample Size" and "Equivalent Degrees of Freedom" Refinements for Inference Using Survey Weights Under Superpopulation Models. *Journal of the American Statistical Association*, Vol. 87, 383-396.

5. Carlson, B.L., A.E. Johnson, and S.B. Cohen (1993). An Evaluation of the Use of Personal Computers for Variance Estimation with Complex Survey Data. *Journal of Official Statistics*, Vol. 9, No. 4, 795-814.

TECHNICAL SUPPLEMENT 8:
DESIGN OF THE HCUP NATIONWIDE INPATIENT SAMPLE, RELEASE 4

INTRODUCTION

The Nationwide Inpatient Sample (NIS) of the Healthcare Cost and Utilization Project (HCUP) was established to provide analyses of hospital utilization across the United States. The NIS, Release 1 covers calendar years 1988-1992. The NIS, Release 2 covers calendar year 1993, the NIS, Release 3 covers calendar year 1994, and the NIS, Release 4 covers calendar year 1995. The target universe includes all acute-care discharges from all community hospitals in the United States; the NIS comprises all discharges from a sample of hospitals in this target universe.

This fourth release of the NIS contains 6.7 million discharges from a sample of 938 hospitals in 19 states. The first release (1988 through 1992) contains 5.2 to 6.2 million discharges per year from a sample of 758 to 875 hospitals per year in 11 states (8 states for 1988). The second release of the NIS contains 6.5 million discharges from a sample of 913 hospitals in 17 states. The third release of the NIS contains 6.4 million discharges from a sample of 904 hospitals in 17 states. Thus, the NIS supports both cross-sectional and longitudinal analyses.

Potential research issues focus on both discharge- and hospital-level outcomes. Discharge outcomes of interest include trends in inpatient treatments with respect to:

- frequency,
- costs,
- lengths of stay,
- effectiveness,
- appropriateness, and
- access to hospital care.

Hospital outcomes of interest include:

- mortality rates,
- complication rates,
- patterns of care,
- diffusion of technology, and
- trends toward specialization.

These and other outcomes are of interest for the nation as a whole and for policy-relevant inpatient subgroups defined by geographic regions, patient demographics, hospital characteristics, physician characteristics, and pay sources.

This report provides a detailed description of the NIS, Release 4 sample design, as well as a summary of the resultant hospital sample. Sample weights were developed to obtain national estimates of hospital and inpatient parameters. These weights and other special-use weights are described in detail. Tables include cumulative information for NIS, Release 1 (1988 through 1992); NIS, Release 2 (1993); NIS, Release 3 (1994); and NIS, Release 4 (1995) to provide a longitudinal view of the database.

THE NIS HOSPITAL UNIVERSE

The hospital universe is defined by all hospitals that were open during any part of the calendar year and were designated as community hospitals in the American Hospital Association (AHA) Annual Survey of Hospitals. For purposes of the NIS, the definition of a community hospital is that used by the AHA: "all nonfederal short-term general and other specialty hospitals, excluding hospital units of institutions." Consequently, Veterans Hospitals and other federal hospitals are excluded. Table 29 shows the number of universe hospitals for each year based on the AHA Annual Survey.

Table 29. Hospital Universe[1]

Year	Number of Hospitals
1988	5,607
1989	5,548
1990	5,468
1991	5,412
1992	5,334
1993	5,313
1994	5,290
1995	5,260

Hospital Merges, Splits, and Closures

All hospital entities that were designated community hospitals in the AHA hospital file were included in the hospital universe. Therefore, if two or more community hospitals merged to create a new community hospital, the original hospitals and the newly-formed hospital were all considered separate hospital entities in the universe for the year of the merge. Likewise, if a community hospital split, the original hospital and all newly created community hospitals were separate entities in the universe for the year of the split. Finally, community hospitals that closed during a year were included as long as they were in operation during some part of the calendar year.

Stratification Variables

To help ensure representativeness, sampling strata were defined based on five hospital characteristics contained in the AHA hospital files. The stratification variables were as follows:

1) *Geographic Region – Northeast, Midwest, West, and South*. This is an important stratifier because practice patterns have been shown to vary substantially by region. For example, lengths of stay tend to be longer in East Coast hospitals than in West Coast hospitals.

2) *Control – government nonfederal, private not-for-profit, and private investor-owned.* These types of hospitals tend to have different missions and different responses to government regulations and policies.

3) *Location – urban or rural.* Government payment policies often differ according to this designation. Also, rural hospitals are generally smaller and offer fewer services than urban hospitals.

4) *Teaching Status – teaching or nonteaching*. The missions of teaching hospitals differ from nonteaching hospitals. In addition, financial considerations differ between these two hospital groups. Currently, the Medicare DRG payments are uniformly higher to teaching hospitals than to nonteaching hospitals. A hospital is considered to be a teaching hospital if it has an AMA-approved residency program or is a member of the Council of Teaching Hospitals (COTH).

5) *Bedsize – small, medium, and large.* Bedsize categories are based on hospital beds, and are specific to the hospital's location and teaching status, as shown in Table 30.

Table 30. Bedsize Categories

Location and Teaching Status	**Hospital Bedsize**		
	Small	**Medium**	**Large**
Rural	1-49	50-99	100+
Urban, nonteaching	1-99	100-199	200+
Urban, teaching	1-299	300-499	500+

Rural hospitals were not split according to teaching status, because rural teaching hospitals were rare. For example, in 1988 there were only 20 rural teaching hospitals. The bedsize categories were defined within location and teaching status because they would otherwise have been redundant. Rural hospitals tend to be small; urban nonteaching hospitals tend to be medium-sized; and urban teaching hospitals tend to be large. Yet it was important to recognize gradations of size within these types of hospitals.

For example, in serving rural discharges, the role of "large" rural hospitals (particularly rural referral centers) often differs from the role of "small" rural hospitals. The cut-off points for the bedsize categories are consistent with those used in *Hospital Statistics,* published annually by the AHA.

To further ensure geographic representativeness, implicit stratification variables included state and three-digit zip code (the first three digits of the hospital's five-digit zip code). The hospitals were sorted according to these variables prior to systematic sampling.

HOSPITAL SAMPLING FRAME

For each year, the *universe* of hospitals was established as all community hospitals located in the U.S. However, it was not feasible to obtain and process all-payer discharge data from a random sample of the entire universe of hospitals for at least two reasons. First, all-payer discharge data were not available from all hospitals for research purposes. Second, based on the experience of prior hospital discharge data collections, it would have been too costly to obtain data from individual hospitals, and it would have been too burdensome to process each hospital's unique data structure.

Therefore, the NIS *sampling frame* was constructed from the subset of universe hospitals that released their discharge data for research use. Two sources for all-payer discharge data were state agencies and private data organizations, primarily state hospital associations. At the time when the sample was drawn, the Agency for Health Care Policy and Research (AHCPR) had agreements with 22 data sources that maintain statewide, all-payer discharge data files to include their data in the HCUP database. However, only 8 states in 1988 and 11 states in 1989-1992 could be included in the first release of the NIS, an additional 6 states have been included in the second and the third release of the NIS, and an additional 2 states have been included in the fourth release of the NIS, as shown in Table 31.

Table 31. States in the Frame for the NIS, Release 1; NIS, Release 2; and NIS, Release 3; and NIS, Release 4

Years	States in the Frame
NIS, Release 1	
1988	California, Colorado, Florida, Iowa, Illinois, Massachusetts, New Jersey, and Washington
1989-1992	Add Arizona, Pennsylvania, and Wisconsin
NIS, Release 2	
1993	Add Connecticut, Kansas, Maryland, New York, Oregon, South Carolina
NIS, Release 3	
1994	No new additions
NIS, Release 4	
1995	Add Missouri, Tennessee

The list of the entire frame of hospitals was composed of all AHA community hospitals in each of the frame states *that could be matched to the discharge data provided to HCUP*, with restrictions on the hospitals that could be included from Illinois, South Carolina, Missouri, and Tennessee. If an AHA community hospital could not be matched to the discharge data provided by the data source, it was eliminated from the sampling frame (but not from the universe).

The Illinois Health Care Cost Containment Council stipulated that no more than 40 percent of the discharges provided by Illinois could be included in the database for any calendar quarter. Consequently, a systematic random sample of Illinois hospitals was drawn for the frame. This prevented the sample from including more than 40 percent of Illinois discharges.

South Carolina and Tennessee stipulated that only hospitals that appear in sampling strata with two or more hospitals were to be included in the NIS. Four South Carolina hospitals were excluded from the frame since there were fewer than 2 South Carolina hospitals in 4 sampling frame strata. The remaining 59 South Carolina community hospitals are included in the frame. Six Tennessee hospitals were excluded from the frame since there were fewer than 2 Tennessee hospitals in 6 sampling frame strata. The remaining 66 Tennessee community hospitals are included in the frame.

Missouri stipulated that only hospitals that had signed releases for public use should be included in the NIS, Release 4. Thirty-five Missouri hospitals signed releases for confidential use only. These hospitals were excluded from the sampling frame, leaving 80 hospitals in the frame.

The number of frame hospitals for each year is shown in Table 32.

Table 32. Hospital Frame

Year	Number of Hospitals
1988	1,247
1989	1,658
1990	1,620
1991	1,604
1992	1,591
1993	2,168
1994	2,135
1995	2,284

HOSPITAL SAMPLE DESIGN

Design Requirements

The NIS is a stratified probability sample of hospitals in the frame, with sampling probabilities calculated to select 20 percent of the universe contained in each stratum. The overall objective was to select a sample of hospitals "generalizable" to the target universe, which includes hospitals outside the frame (zero probability of selection). Moreover, this sample was to be geographically dispersed, yet drawn from the subset of states with inpatient discharge data that agreed to provide such data to the project.

It should be possible, for example, to estimate DRG-specific average lengths of stay over all U.S. hospitals using weighted average lengths of stay, based on averages or regression estimates from the NIS. Ideally, relationships among outcomes and their correlates estimated from the NIS should generally hold across all U.S. hospitals. However, since only 19 states contributed data to this fourth release, some estimates may differ from estimates from comparative data sources. When possible, estimates based on the NIS should be checked against national benchmarks, such as Medicare data or data from the National Hospital Discharge Survey to determine the appropriateness of the NIS for specific analyses.

The target sample size was 20 percent of the total number of community hospitals in the U.S. for 1995. This sample size was determined by AHCPR based on their experience with similar research databases.

Alternative stratified sampling allocation schemes were considered. However, allocation proportional to the number of hospitals is preferred for several reasons:

- Fewer than 10 percent of government-planned database applications will produce nationwide estimates. The major government applications will investigate relationships among variables. For example, government researchers will do a substantial amount of regression modeling with these data.

- The HCUP-2 sample[2] used the same stratification and allocation scheme, and it has served AHCPR analysts well. Moreover, the large number of sample hospitals and discharges seemingly reduced the need for variance-reducing allocation schemes.

- AHCPR researchers wanted a simple, easily understood sampling methodology. It was an appealing idea that the NIS sample could be a "miniaturization" of the universe of hospitals (with the obvious geographical limitations imposed by data availability).

- AHCPR statisticians considered other optimal allocation schemes, including sampling hospitals with probabilities proportional to size (number of discharges), and they concluded that sampling with probability proportional to the number of hospitals was preferable. Even though it was recognized that the approach chosen would not be as efficient, the extremely large sample sizes yield good estimates. Furthermore, because the data are to be used for purposes other than producing national estimates, it is critical that all hospital types (including small hospitals) are adequately represented.

Hospital Sampling Procedure

Once the universe of hospitals was stratified, up to 20 percent of the total number of U.S. hospitals was randomly selected within each stratum. If too few frame hospitals were in the stratum, then all frame hospitals were selected for the NIS, subject to sampling restrictions specified by states. To simplify variance calculations, at least two hospitals were drawn from each stratum. If fewer than two frame hospitals were contained in a stratum, then that stratum was merged with an "adjacent" stratum containing hospitals with similar characteristics.

A systematic random sample was drawn from each stratum, after sorting hospitals by state within each stratum, then by the three-digit zip code (the first three digits of the hospital's five-digit zip code) within each state, and then by a random number within each three-digit zip code. These sorts ensured further geographic generalizability of hospitals within the frame states, and random ordering of hospitals within three-digit zip codes.

Generally, three-digit zip codes that are near in value are geographically near within a state. Furthermore, the U.S. Postal Service locates regional mail distribution centers at the three-digit level. Thus, the boundaries tend to be a compromise between geographic size and population size.

1995 NIS Hospital Sampling Procedure

The 1995 sample was drawn by a procedure that retained most of the 1994 hospitals, while allowing hospitals new to the frame an opportunity to enter the 1995 NIS.

Even in frame states that were present in the 1994 sample, hospitals that opened in 1995 needed a chance to enter the sample. Also, hospitals that changed strata between 1994 and 1995 were considered new to the 1995 frame.

Consequently, a recursive procedure was developed to update the sample from year to year in a way that properly accounted for changes in stratum size, composition, and sampling rate. The goal of this procedure was to maximize the year-to-year overlap among sample hospitals, yet keep the sampling rate constant for all hospitals *within a stratum*.

The following procedure provides rules for creating a "year 2" sample, given that a "year 1" sample had already been drawn. In this example, year 1 would be 1994 and year 2 would be 1995. All notation is assumed to refer to sizes and probabilities within a particular stratum.

Probabilities P_1 and P_2 were calculated for sampling hospitals from the frame within the stratum for year 1 and year 2, respectively, based on the frame and universe for year 1 and year 2, respectively. These probabilities were set by the same algorithm used to calculate P for the 1988 hospital sample (see Technical Supplement: *Design of the HCUP Nationwide Inpatient Sample, Release 1*, section "1988 NIS Hospital Sampling Procedure.")

Now consider the three possibilities associated with changes between years 1 and 2 in the stratum-specific hospital sampling probabilities:

1. $P_2 = P_1$: The target probability was unchanged.

2. $P_2 < P_1$: The target probability decreased.

3. $P_2 > P_1$: The target probability increased.

Below is the procedure used for each of these three cases with one exception: if the stratum-specific probability of selection P_2 was equal to 1, then all frame hospitals were selected for the year 2 sample, regardless of the value of P_1.

Stratum-Specific Sampling Rates the Same ($P_2 = P_1$). If the probability P_2 was the same as P_1, all hospitals in the year 1 sample that remained in the year 2 frame were retained for the year 2 sample. Any new frame hospitals (those in the year 2 frame but not in the year 1 frame) were selected at the rate P_2, using the systematic sampling method described for the 1988 sample selection in Technical Supplement: *Design of the HCUP Nationwide Inpatient Sample, Release 1*.

Stratum-Specific Sampling Rate Decreased ($P_2 < P_1$). Now consider the case where the probability of selection decreased between years 1 and 2. First, hospitals new to the frame were sampled with probability P_2. Second, hospitals previously selected for the year 1 sample (that remained in the year 2 frame) were selected for the year 2 sample with probability $P_2 \div P_1$.

The justification for this second procedure was straightforward. For the year 1 sample hospitals that stayed in the frame, the year 1 sample was viewed as the first stage of a two-stage sampling process. The first stage was carried out at the sampling rate of P_1. The second stage was carried out at the sampling rate of $P_2 \div P_1$. Consequently, the "overall" probability of selection was $P_1 \times P_2 \div P_1 = P_2$.

Stratum-Specific Sampling Rate Increased ($P_2 > P_1$). The procedures associated with the case in which the probability of selection was increased between year 1 and year 2 were equally straightforward. First, hospitals new to the frame were sampled with probability P_2. Second, hospitals that were selected in year 1 (that remained in the year 2 frame) were selected for the year 2 sample. Third, hospitals that were in the frame for both years 1 and 2, but not selected for the year 1 sample, were selected for the year 2 sample with probability $(P_2\text{-}P_1) \div (1\text{-}P_1)$.

The justification for this sampling rate, $(P_2\text{-}P_1) \div (1\text{-}P_1)$, is somewhat complex. In year 1 certain frame hospitals were included in the sample at the rate P_1. This can also be viewed as having excluded a set of hospitals at the rate $(1\text{-}P_1)$. Likewise, in year 2 it was imperative that each hospital excluded from the year 1 sample be excluded from the year 2 sample at an overall rate of $(1\text{-}P_2)$.

Since $P_2 > P_1$, then $(1\text{-}P_2) < (1\text{-}P_1)$. Therefore, just as was done for the case of $P_2 < P_1$, multistage selection was implemented. However, it was implemented for exclusion rather than inclusion.

Therefore, those hospitals excluded from the year 1 sample were also excluded from the year 2 sample at the rate $S = (1\text{-}P_2) \div (1\text{-}P_1)$. This gave them the desired overall *exclusion* rate of $(1\text{-}P_1) \times (1\text{-}P_2) \div (1\text{-}P_1) = (1\text{-}P_2)$. Consequently, the *inclusion* rate for these hospitals was set at $1\text{-}S = (P_2\text{-}P_1) \div (1\text{-}P_1)$.

Zero-Weight Hospitals

The 1995 sample contains no zero-weight hospitals. For a description of zero-weight hospitals in the 1988-1992 sample, see the Technical Supplement: *Design of the HCUP Nationwide Inpatient Sample, Release 1*.

Ten Percent Subsamples

Two nonoverlapping 10 percent subsamples of discharges were drawn from the NIS file for each year. The subsamples were selected by drawing every tenth discharge starting with two different starting points (randomly selected between 1 and 10). Having a different starting point for each of the two subsamples guaranteed that they would not overlap. Discharges were sampled so that 10 percent of each hospital's discharges in each quarter were selected for each of the subsamples. The two samples can be combined to form a single, generalizable 20 percent subsample of discharges.

FINAL HOSPITAL SAMPLE

The annual numbers of hospitals and discharges in NIS, Release 1; NIS, Release 2; NIS, Release 3; and NIS, Release 4 are shown in Table 33, for both the regular NIS sample and the total sample (which includes zero-weight hospitals for 1988-1992).

Table 33. NIS Hospital Sample

	Regular Sample		Total Sample	
Year	**Number of Hospitals**	**Number of Discharges**	**Number of Hospitals**	**Number of Discharges**
NIS, Release 1				
1988	758	5,242,904	759	5,265,756
1989	875	6,067,667	882	6,110,064
1990	861	6,156,638	871	6,268,515
1991	847	5,984,270	859	6,156,188
1992	838	6,008,001	856	6,195,744
NIS, Release 2				
1993	913	6,538,976	913	6,538,976
NIS, Release 3				
1994	904	6,385,011	904	6,385,011
NIS, Release 4				
1995	938	6,714,935	938	6,714,935
Total		49,098,402		49,635,189

A more detailed breakdown of the 1995 NIS hospital sample by geographic region is shown in Table 34. For each geographic region, Table 34 shows the number of:

- universe hospitals (Universe),
- frame hospitals (Frame),
- sampled hospitals (Sample),
- target hospitals (Target = 20 percent of the universe), and
- shortfall hospitals (Shortfall = Sample - Target).

Table 34. Number of Hospitals in Universe, Frame, Regular Sample, Target, and Shortfall By Region, 1995

Region	Universe	Frame	Sample	Target	Shortfall
NE	772	648	162	154	8
MW	1,507	557	317	302	15
S	2,004	375	278	401	-123
W	977	704	181	195	-14
Total	5,260	2,284	938	1,052	-114

For example, in 1995 the Northeast region contained 772 hospitals in the universe. It also contained 648 hospitals in the frame, of which 162 hospitals were drawn for the sample. This was 8 hospitals more than the target sample size of 154.

Table 35 shows the number of hospitals in the universe, frame, and regular sample for each state in the sampling frame for 1995. In all states except Illinois, South Carolina, Tennessee, and Missouri, the difference between the universe and the frame represents the difference in the number of community hospitals in the 1995 AHA Annual Survey of Hospitals and the number of community hospitals for which data were supplied to HCUP. As explained earlier, the number of hospitals in the Illinois frame is approximately 53 percent of the hospitals in the Illinois universe in order to comply with the agreement with the data source concerning the restriction on the number of Illinois discharges. The number of hospitals in the South Carolina frame is 7 fewer than the South Carolina universe. Four hospitals were excluded because of sampling restrictions stipulated by South Carolina, and 3 hospitals were not included in the data supplied to HCUP. The number of hospitals in the Tennessee frame is 63 fewer than the Tennessee universe. Six hospitals were excluded because of sampling restrictions stipulated by Tennessee, and 57 hospitals were not included in the data supplied to HCUP. The number of hospitals in the Missouri frame is 49 fewer than the Missouri universe. Thirty-five hospitals were excluded because they signed release for confidential use only, and 14 hospitals were not included in the data supplied to HCUP.

The number of hospitals in the NIS hospital samples that continue across multiple sample years is shown in Table 36. This table will be of interest to those who may combine Releases 1, 2, 3, and 4 of the NIS. Table 36 shows that longitudinal cohorts that span several years and include 1988 and 1993 are the lowest in number of continuing sample hospitals. For example, if 1988 is taken as a starting year, only 38.1 percent of the 1988 hospital sample continued in the 1995 sample (289 of 758).

Table 35. Number of Hospitals in the Universe, Frame, and Regular Sample for States in the Sampling Frame: 1995

State	Universe	Frame	Sample
AZ	62	60	15
CA	426	425	105
CO	70	69	22
CT	34	32	9
FL	215	200	141
IA	116	116	54
IL	207	110	73
KS	134	124	61
MA	96	82	25
MD	50	50	39
MO	129	80	49
NJ	92	85	18
NY	231	230	59
OR	64	62	17
PA	225	219	51
SC	66	59	46
TN	129	66	52
WA	89	88	22
WI	127	127	80
Total	2562	2284	938

Table 36. Number of Hospitals and Discharges in Longitudinal Cohort

Number of Years	Calendar Years	Longitudinal Regular Sample Hospitals	% of Base Year Sample	Longitudinal Regular Sample Discharges
2	1988-1989	610	80.5	8,492,039
	1989-1990	815	93.1	11,525,749
	1990-1991	802	93.1	11,297,175
	1991-1992	781	92.2	11,272,981
	1992-1993	609	72.7	8,804,638
	1993-1994	693	75.9	10,271,404
	1994-1995	762	84.3	10,747,682
3	1988-1990	573	75.6	12,168,677
	1989-1991	763	87.2	16,074,381
	1990-1992	745	86.5	16,085,651
	1991-1993	570	67.3	12,559,421
	1992-1994	540	64.4	11,279,667
	1993-1995	598	65.5	13,241,070
4	1988-1991	542	71.5	15,096,807
	1989-1992	709	81.0	20,340,970
	1990-1993	548	63.6	16,023,500
	1991-1994	508	60.0	14,481,319
	1992-1995	464	55.4	12,712,613
5	1988-1992	502	66.2	18,106,098
	1989-1993	523	59.8	19,000,777
	1990-1994	490	56.9	17,437,229
	1991-1995	439	51.8	15,405,253
6	1988-1993	378	49.9	16,906,818
	1989-1994	471	53.8	19,987,910
	1990-1995	422	49.0	14,817,797
7	1988-1994	335	44.2	17,128,064
	1989-1995	408	46.6	19,924,107
8	1988-1995	289	38.1	16,658,485

SAMPLING WEIGHTS

Although the sampling design was simple and straightforward, it is necessary to incorporate sample weights to obtain state and national estimates. Therefore, sample weights were developed separately for hospital- and discharge-level analyses. Three hospital-level weights were developed to weight NIS sample hospitals to the state, frame, and universe. Similarly, three discharge-level weights were developed to weight NIS sample discharges to the state, frame, and universe.

Hospital-Level Sampling Weights

Universe Hospital Weights. Hospital weights to the universe were calculated by post-stratification. For each year, hospitals were stratified on the same variables that were used for sampling: geographic region, urban/rural location, teaching status, bedsize, and control. The strata that were collapsed for sampling were also collapsed for sample weight calculations. Within stratum s, each NIS sample hospital's universe weight was calculated as:

$$W_s(\text{universe}) = N_s(\text{universe}) \div N_s(\text{sample}),$$

where N_s(universe) and N_s(sample) were the number of community hospitals within stratum s in the universe and sample, respectively. Thus, each hospital's universe weight is equal to the number of universe hospitals it represented during that year.

Frame Hospital Weights. Hospital-level sampling weights were also calculated to represent the entire collection of states in the frame using the same post-stratification scheme as described above for the weights to represent the universe. For each year, within stratum s, each NIS sample hospital's frame weight was calculated as:

$$W_s(\text{frame}) = N_s(\text{frame}) \div N_s(\text{sample}).$$

N_s(frame) was the total number of universe community hospitals within stratum s in the states that contributed data to the frame. N_s(sample) was the number of sample hospitals selected for the NIS in stratum s. Thus, each hospital's frame weight is equal to the number of universe hospitals it represented in the frame states during that year.

State Hospital Weights. For each year, a hospital's weight to its state was calculated in a similar fashion. Within each state, strata often had to be collapsed after sample selection for development of weights to ensure a minimum of two sample hospitals within each stratum. For each state and each year, within stratum s, each NIS sample hospital's state weight was calculated as:

$$W_s(\text{state}) = N_s(\text{state}) \div N_s(\text{state sample}).$$

N_s(state) was the number of universe community hospitals in the state within stratum s. N_s(state sample) was the number of hospitals selected for the NIS from that state in stratum s. Thus, each hospital's state weight is equal to the number of hospitals that it represented in its state during that year.

All of these hospital weights can be rescaled if necessary for selected analyses, to sum to the NIS hospital sample size each year.

Discharge-Level Sampling Weights

The calculations for discharge-level sampling weights were very similar to the calculations of hospital-level sampling weights. The discharge weights usually are constant for all discharges within a stratum.

The only exceptions were for strata with sample hospitals that, according to the AHA files, were open for the entire year but contributed less than their full year of data to the NIS. For those hospitals, we *adjusted* the number of observed discharges by a factor 4 ÷ Q, where Q was the number of calendar quarters that the hospital contributed discharges to the NIS. For example, when a sample hospital contributed only two quarters of discharge data to the NIS, the *adjusted* number of discharges was double the observed number.

With that minor adjustment, each discharge weight is essentially equal to the number of reference (universe, frame, or state) discharges that each sampled discharge represented in its stratum. This calculation was possible because the number of total discharges was available for every hospital in the universe from the AHA files. Each universe hospital's AHA discharge total was calculated as the sum of newborns and total facility discharges.

Universe Discharge Weights. Discharge weights to the universe were calculated by post-stratification. Hospitals were stratified just as they were for universe hospital weight calculations. Within stratum s, for hospital i, each NIS sample discharge's universe weight was calculated as:

$$DW_{is}(\text{universe}) = [DN_s(\text{universe}) \div ADN_s(\text{sample})] * (4 \div Q_i),$$

where DN_s(universe) was the number of discharges from community hospitals in the universe within stratum s; ADN_s(sample) was the number of *adjusted* discharges from sample hospitals selected for the NIS; and Q_i was the number of quarters of discharge data contributed by hospital i to the NIS (usually $Q_i = 4$). Thus, each discharge's weight is equal to the number of universe discharges it represented in stratum s during that year.

Frame Discharge Weights. Discharge-level sampling weights were also calculated to represent all discharges from the entire collection of states in the frame using the same post-stratification scheme described above for the discharge weights to represent the universe. For each year, within stratum s, for hospital i, each NIS sample discharge's frame weight was calculated as:

$$W_{is}(\text{frame}) = [DN_s(\text{frame}) \div ADN_s(\text{sample})] * (4 \div Q_i),$$

DN_s(frame) was the number of discharges from all community hospitals in the states that contributed to the frame within stratum s. ADN_s(sample) was the number of *adjusted* discharges from sample hospitals selected for the NIS in stratum s. Q_i was the number of quarters of discharge data contributed by hospital i to the NIS (usually $Q_i = 4$). Thus, each discharge's frame weight is equal to the number of discharges it represented in the frame states during that year.

State Discharge Weights. A discharge's weight to its state was similarly calculated. Strata were collapsed in the same way as they were for the state hospital weights to ensure a minimum of two sample hospitals within each stratum. Within stratum s, for hospital i, each NIS sample discharge's state weight was calculated as:

$$W_{is}(\text{state}) = [DN_s(\text{state}) \div ADN_s(\text{state sample})] * (4 \div Q_i),$$

DN_s(state) was the number of discharges from all community hospitals in the state within stratum s. ADN_s(state sample) was the *adjusted* number of discharges from hospitals selected for the NIS from that state in stratum s. Q_i was the number of quarters of discharge data contributed by hospital i to the NIS (usually $Q_i = 4$). Thus, each discharge's state weight is equal to the number of discharges that it represented in its state during that year.

All of these discharge weights can be rescaled if necessary for selected analyses, to sum to the NIS discharge sample size each year.

Discharge Weights for 10 Percent Subsamples

In the 10 percent subsamples, each discharge had a 10 percent chance of being drawn. Therefore, the discharge weights contained in the Hospital Weights file can be multiplied by 10 for each of the subsamples, or multiplied by 5 for the two subsamples combined.

DATA ANALYSIS

Variance Calculations

It may be important for researchers to calculate a measure of precision for some estimates based on the NIS sample data. Variance estimates must take into account both the sampling design and the form of the statistic. The sampling design was a stratified, single-stage cluster sample. A stratified random sample of hospitals (clusters) were drawn and then *all* discharges were included from each selected hospital.

If hospitals inside the frame were similar to hospitals outside the frame, the sample hospitals can be treated as if they were randomly selected from the entire universe of hospitals within each stratum. Standard formulas for a stratified, single-stage cluster sampling without replacement could be used to calculate statistics and their variances in most applications.

A multitude of statistics can be estimated from the NIS data. Several computer programs are listed below that calculate statistics and their variances from sample survey data. Some of these programs use general methods of variance calculations (e.g., the jackknife and balanced half-sample replications) that take into account the sampling design. However, it may be desirable to calculate variances using formulas specifically developed for some statistics.

In most cases, computer programs are readily available to perform these calculations. For instance, OSIRIS IV, developed at the University of Michigan, and SUDAAN, developed at the Research Triangle Institute, do calculations for numerous statistics arising from the stratified, single-stage cluster sampling design. An example of using SUDAAN to calculate variances in the NIS is presented in Technical Supplement: *Calculating Variances Using Data from the HCUP Nationwide Inpatient Sample*.[3]

These variance calculations are based on finite-sample theory, which is an appropriate method for obtaining cross-sectional, nationwide estimates of outcomes. According to finite-sample theory, the intent of the estimation process is to obtain estimates that are precise representations of the nationwide population at a specific point in time. In the context of the NIS, any estimates that attempt to accurately describe characteristics (such as expenditure and utilization patterns or hospital market factors) and interrelationships among characteristics of hospitals and discharges during a specific year from 1988 to 1995 should be governed by finite-sample theory.

Alternatively, in the study of hypothetical population outcomes not limited to a specific point in time, analysts may be less interested in specific characteristics from the finite population (and time period) from which the *sample* was drawn, than they are in hypothetical characteristics of a conceptual "superpopulation" from which any particular finite *population* in a given year might have been drawn. According to this superpopulation model, the nationwide population in a given year is only a snapshot in time of the possible interrelationships among hospital, market, and discharge characteristics. In a given year, all possible interactions between such characteristics may not have been observed, but analysts may wish to predict or simulate interrelationships that may occur in the future.

Under the finite-population model, the variances of estimates approach zero as the sampling fraction approaches one, since the population is defined at that point in time, and because the estimate is for a characteristic as it existed at the time of sampling. This is in contrast to the superpopulation model, which adopts a stochastic viewpoint rather than a deterministic viewpoint. That is, the nationwide population in a particular year is viewed as a random sample of some underlying superpopulation over time.

Different methods are used for calculating variances under the two sample theories. Under the superpopulation (stochastic) model, procedures (such as those described by Potthoff, Woodbury, and Manton[4]) have been developed to draw inferences using weights from complex samples. In this context, the survey weights are not used to weight the sampled cases to the universe, because the universe is conceptually infinite in size. Instead, these weights are used to produce unbiased estimates of parameters that govern the superpopulation.

In summary, the choice of an appropriate method for calculating variances for nationwide estimates depends on the type of measure and the intent of the estimation process.

Computer Software for Variance Calculations

The hospital weights will be useful for producing hospital-level statistics for analyses that use the *hospital* as the unit of analysis, and the discharge weights will be useful for producing discharge-level statistics for analyses that use the *discharge* as the unit of analysis. These would be used to weight the sample data in estimating population statistics.

Several statistical programming packages allow weighted analyses.[5] For example, nearly all SAS (Statistical Analysis System) procedures incorporate weights.

In addition, several publicly available subroutines have been developed specifically for calculating statistics and their standard errors from survey data:

- OSIRIS IV was developed by L. Kish, N. Van Eck, and M. Frankel at the Survey Research Center, University of Michigan. It consists of two main programs for estimating variances from complex survey designs.

- SUDAAN, a set of SAS subroutines, was developed at the Research Triangle Institute by B. V. Shah. It is adequate for handling most survey designs with stratification. The procedures can handle estimation and variance estimation for means, proportions, ratios, and regression coefficients.

- SUPER CARP (Cluster Analysis and Regression Program) was developed at Iowa State University by W. Fuller, M. Hidiroglou, and R. Hickman. This program computes estimates and variance estimates for multistage, stratified sampling designs with arbitrary probabilities of selection. It can handle estimated totals, means, ratios, and regression estimates.

The NIS database includes a Hospital Weights file with variables required by these programs to calculate finite population statistics. In addition to the sample weights described earlier, hospital identifiers (PSUs), stratification variables, and stratum-specific totals for the numbers of discharges and hospitals are included so that finite-population corrections (FPCs) can be applied to variance estimates.

In addition to these subroutines, standard errors can be estimated by validation and cross-validation techniques. Given that a very large number of observations will be available for most analyses, it may be feasible to set aside a part of the data for validation purposes. Standard errors and confidence intervals can then be calculated from the validation data. If the analytical file is too small to set aside a large validation sample, cross-validation techniques may be used.

For example, tenfold cross-validation would split the data into ten equal-sized subsets. The estimation would take place in ten iterations. At each iteration, the outcome of interest is predicted for one-tenth of the observations by an estimate based on a model fit to the other nine-tenths of the observations. Unbiased estimates of error variance are then obtained by comparing the actual values to the predicted values obtained in this manner.

Finally, it should be noted that a large array of hospital-level variables are available for the entire universe of hospitals, including those outside the sampling frame. For instance, the variables from the AHA surveys and from the Medicare Cost Reports are available for nearly all hospitals. To the extent that hospital-level outcomes correlate with these variables, they may be used to sharpen regional and nationwide estimates.

As a simple example, each hospital's number of C-sections would be correlated with their total number of deliveries. The number of C-sections must be obtained from discharge data, but the number of deliveries is available from AHA data. Thus, if a regression can be fit predicting C-sections from deliveries based on the NIS data, that regression can then be used to obtain hospital-specific estimates of the number of C-sections for all hospitals in the universe.

Longitudinal Analyses

As previously shown in Table 36, hospitals that continue in the NIS for multiple consecutive years are a subset of the hospitals in the NIS for any one of those years. Consequently, longitudinal analyses of hospital-level outcomes may be biased if they are based on any subset of NIS hospitals limited to continuous NIS membership. In particular, such subsets would tend to contain fewer hospitals that opened, closed, split, merged, or changed strata. Further, the sample weights were developed as annual, cross-sectional weights rather than longitudinal weights. Therefore, different weights might be required, depending on the statistical methods employed by the analyst.

One approach to consider in hospital-level longitudinal analyses is to use repeated-measure models that allow hospitals to have missing values for some years. However, the data are not actually missing for some hospitals, such as those that closed during the study period. In any case, the analyses may be more efficient (e.g., produce more precise estimates) if they account for the potential correlation between repeated measures on the same hospital over time, yet incorporate data from all hospitals in the sample during the study period.

Discharge Subsamples

The two nonoverlapping 10 percent subsamples of discharges were drawn from the NIS file for each year for several reasons pertaining to data analysis. One reason for creating the subsamples was to reduce processing costs for selected studies that will not require the entire NIS. Another reason is that the two subsamples may be used to validate models and obtain unbiased estimates of standard errors. That is, one subsample may be used to estimate statistical models, and the other subsample may be used to test the fit of those models on new data. This is a very important analytical step, particularly in exploratory studies, where one runs the risk of fitting noise.

For example, it is well known that the percentage of variance explained by a regression, R^2, is generally overestimated by the data used to fit a model. The regression model could be estimated from the first subsample and then applied to the second subsample. The squared correlation between the actual and predicted value in the second subsample is an unbiased estimate of the model's true explanatory power when applied to new data.

ENDNOTES

1. Most AHA surveys do not cover a January-to-December calendar year. The number of hospitals for 1988-1991 are based on the HCUP calendar-year version of the AHA Annual Survey files. To create a calendar-year reporting period, data from the AHA surveys must be apportioned in some manner across calendar years. Survey responses were converted to calendar-year periods for 1988-1991 by merging data from adjacent survey years. The number of hospitals for 1992-1994 are based on the AHA Annual Survey files.

2. Coffey, R. and D. Farley (1988, July). *HCUP-2 Project Overview,* (DHHS Publication No. (PHS) 88-3428. Hospital Studies Program Research Note 10, National Center for Health Services Research and Health Care Technology Assessment, Rockville, MD: Public Health Service.

3. Duffy, S.Q. and J.P. Sommers (1996, March). *Calculating Variances Using Data from the HCUP Nationwide Inpatient Sample*. Rockville, MD: Agency for Health Care Policy and Research.

4. Potthoff, R.F., M.A. Woodbury, and K.G. Manton (1992). "Equivalent Sample Size" and "Equivalent Degrees of Freedom" Refinements for Inference Using Survey Weights Under Superpopulation Models. *Journal of the American Statistical Association*, Vol. 87, 383-396.

5. Carlson, B.L., A.E. Johnson, and S.B. Cohen (1993). An Evaluation of the Use of Personal Computers for Variance Estimation with Complex Survey Data. *Journal of Official Statistics*, Vol. 9, No. 4, 795-814.

TECHNICAL SUPPLEMENT 9:
DESIGN OF THE HCUP NATIONWIDE INPATIENT SAMPLE, RELEASE 5

INTRODUCTION

The Nationwide Inpatient Sample (NIS) of the Healthcare Cost and Utilization Project (HCUP) was established to provide analyses of hospital utilization across the United States. The target universe includes all acute-care discharges from all community hospitals in the United States; the NIS comprises all discharges from a sample of hospitals in this target universe.

NIS Release	Calendar Year	States	Sample Hospitals	Sample Discharge (millions)
1	1988–1992	8–11	758–875	5.2–6.2
2	1993	17	913	6.5
3	1994	17	904	6.4
4	1995	19	938	6.7
5	1996	19	906	6.5

Thus, the NIS supports both cross-sectional and longitudinal analyses.

Potential research issues focus on both discharge- and hospital-level outcomes. Discharge outcomes of interest include trends in inpatient treatments with respect to:

- frequency,
- costs,
- lengths of stay,
- effectiveness,
- appropriateness, and
- access to hospital care.

Hospital outcomes of interest include:

- mortality rates,
- complication rates,
- patterns of care,
- diffusion of technology, and
- trends toward specialization.

These and other outcomes are of interest for the nation as a whole and for policy-relevant inpatient subgroups defined by geographic regions, patient demographics, hospital characteristics, physician characteristics, and pay sources.

This report provides a detailed description of the NIS, Release 5 sample design, as well as a summary of the resultant hospital sample. Sample weights were developed to obtain national estimates of hospital and inpatient parameters. These weights and other special-use weights are described in detail. Tables include cumulative information for all five NIS Releases to provide a longitudinal view of the database.

THE NIS HOSPITAL UNIVERSE

The hospital universe is defined by all hospitals that were open during any part of the calendar year and were designated as community hospitals in the American Hospital Association (AHA) Annual Survey of Hospitals. For purposes of the NIS, the definition of a community hospital is that used by the AHA: "all nonfederal short-term general and other specialty hospitals, excluding hospital units of institutions." Consequently, Veterans Hospitals and other federal hospitals are excluded. Table 37 shows the number of universe hospitals for each year based on the AHA Annual Survey.

Table 37. Hospital Universe[1]

Year	Number of Hospitals
1988	5,607
1989	5,548
1990	5,468
1991	5,412
1992	5,334
1993	5,313
1994	5,290
1995	5,260
1996	5,182

Hospital Merges, Splits, and Closures

All hospital entities that were designated community hospitals in the AHA hospital file were included in the hospital universe. Therefore, if two or more community hospitals merged to create a new community hospital, the original hospitals and the newly-formed hospital were all considered separate hospital entities in the universe for the year of the merge. Likewise, if a community hospital split, the original hospital and all newly created community hospitals were separate entities in the universe for the year of the split. Finally, community hospitals that closed during a year were included as long as they were in operation during some part of the calendar year.

Stratification Variables

To help ensure representativeness, sampling strata were defined based on five hospital characteristics contained in the AHA hospital files. The stratification variables were as follows:

1) *Geographic Region – Northeast, Midwest, West, and South*. This is an important stratifier because practice patterns have been shown to vary substantially by region. For example, lengths of stay tend to be longer in East Coast hospitals than in West Coast hospitals.

2) *Control – government nonfederal, private not-for-profit, and private investor-owned.* These types of hospitals tend to have different missions and different responses to government regulations and policies.

3) *Location – urban or rural.* Government payment policies often differ according to this designation. Also, rural hospitals are generally smaller and offer fewer services than urban hospitals.

4) *Teaching Status – teaching or nonteaching*. The missions of teaching hospitals differ from nonteaching hospitals. In addition, financial considerations differ between these two hospital groups. Currently, the Medicare DRG payments are uniformly higher to teaching hospitals than to nonteaching hospitals. A hospital is considered to be a teaching hospital if it has an AMA-approved residency program or is a member of the Council of Teaching Hospitals (COTH).

5) *Bedsize – small, medium, and large.* Bedsize categories are based on hospital beds, and are specific to the hospital's location and teaching status, as shown in Table 38.

Table 38. Bedsize Categories

Location and Teaching Status	**Hospital Bedsize**		
	Small	**Medium**	**Large**
Rural	1-49	50-99	100+
Urban, nonteaching	1-99	100-199	200+
Urban, teaching	1-299	300-499	500+

Rural hospitals were not split according to teaching status, because rural teaching hospitals were rare. For example, in 1988 there were only 20 rural teaching hospitals. The bedsize categories were defined within location and teaching status because they would otherwise have been redundant. Rural hospitals tend to be small; urban nonteaching hospitals tend to be medium-sized; and urban teaching hospitals tend to be large. Yet it was important to recognize gradations of size within these types of hospitals.

For example, in serving rural discharges, the role of "large" rural hospitals (particularly rural referral centers) often differs from the role of "small" rural hospitals. The cut-off points for the bedsize categories are consistent with those used in *Hospital Statistics,* published annually by the AHA.

To further ensure geographic representativeness, implicit stratification variables included state and three-digit zip code (the first three digits of the hospital's five-digit zip code). The hospitals were sorted according to these variables prior to systematic sampling.

HOSPITAL SAMPLING FRAME

For each year, the *universe* of hospitals was established as all community hospitals located in the U.S. However, it was not feasible to obtain and process all-payer discharge data from a random sample of the entire universe of hospitals for at least two reasons. First, all-payer discharge data were not available from all hospitals for research purposes. Second, based on the experience of prior hospital discharge data collections, it would have been too costly to obtain data from individual hospitals, and it would have been too burdensome to process each hospital's unique data structure.

Therefore, the NIS *sampling frame* was constructed from the subset of universe hospitals that released their discharge data for research use. Two sources for all-payer discharge data were state agencies and private data organizations, primarily state hospital associations. At the time when the sample was drawn, the Agency for Health Care Policy and Research (AHCPR) had agreements with 22 data sources that maintain statewide, all-payer discharge data files to include their data in the HCUP database. However, only 8 states in 1988 and 11 states in 1989-1992 could be included in the first release of the NIS, an additional 6 states were included in the second and the third release of the NIS, and another 2 states were included in the fourth and the fifth release of the NIS, as shown in Table 39.

Table 39. States in the Frame for NIS Releases

Years	States in the Frame
NIS, Release 1	
1988	California, Colorado, Florida, Iowa, Illinois, Massachusetts, New Jersey, and Washington
1989-1992	Add Arizona, Pennsylvania, and Wisconsin
NIS, Release 2	
1993	Add Connecticut, Kansas, Maryland, New York, Oregon, South Carolina
NIS, Release 3	
1994	No new additions
NIS, Release 4	
1995	Add Missouri, Tennessee
NIS, Release 5	
1996	No new additions

The list of the entire frame of hospitals was composed of all AHA community hospitals in each of the frame states *that could be matched to the discharge data provided to HCUP*. If an AHA community hospital could not be matched to the discharge data provided by the data source, it was eliminated from the sampling frame (but not from the universe). Further restrictions were put on the sampling frames for Illinois, South Carolina, Missouri, and Tennessee.

The Illinois Health Care Cost Containment Council stipulated that no more than 40 percent of the discharges provided by Illinois could be included in the database for any calendar quarter. Consequently, a systematic random sample of Illinois hospitals was drawn for the 1996 frame. This prevented the sample from including more than 40 percent of Illinois discharges.

South Carolina and Tennessee stipulated that only hospitals that appear in sampling strata with two or more hospitals were to be included in the NIS. Six South Carolina hospitals were excluded from the 1996 frame since there was only one South Carolina hospital in six sampling frame strata. The remaining 55 South Carolina community hospitals in 1996 were included in the frame. Four Tennessee hospitals were excluded from the 1996 frame since there was only one Tennessee hospital in four sampling frame strata. The remaining 88 Tennessee community hospitals in 1996 were included in the frame.

Missouri stipulated that only hospitals that had signed releases for public use should be included in the NIS, Release 5. For 1996, thirty-five Missouri hospitals signed releases for confidential use only. These hospitals were excluded from the sampling frame, leaving 79 hospitals in the 1996 frame.

The number of frame hospitals for each year is shown in Table 40.

Table 40. Hospital Frame

Year	Number of Hospitals
1988	1,247
1989	1,658
1990	1,620
1991	1,604
1992	1,591
1993	2,168
1994	2,135
1995	2,284
1996	2,268

HOSPITAL SAMPLE DESIGN

Design Requirements

The NIS is a stratified probability sample of hospitals in the frame, with sampling probabilities calculated to select 20 percent of the universe contained in each stratum. The overall objective was to select a sample of hospitals "generalizable" to the target universe, which includes hospitals outside the frame (zero probability of selection). Moreover, this sample was to be geographically dispersed, yet drawn from the subset of states with inpatient discharge data that agreed to provide such data to the project.

It should be possible, for example, to estimate DRG-specific average lengths of stay over all U.S. hospitals using weighted average lengths of stay, based on averages or regression estimates from the NIS. Ideally, relationships among outcomes and their correlates estimated from the NIS should generally hold across all U.S. hospitals. However, since only 19 states contributed data to this fifth release, some estimates may differ from estimates from comparative data sources. When possible, estimates based on the NIS should be checked against national benchmarks, such as Medicare data or data from the National Hospital Discharge Survey to determine the appropriateness of the NIS for specific analyses (see the Technical Supplement: *Comparative Analysis of HCUP and NHDS Inpatient Discharge Data*).

The target sample size was 20 percent of the total number of community hospitals in the U.S. for 1996. This sample size was determined by AHCPR based on their experience with similar research databases.

Alternative stratified sampling allocation schemes were considered. However, allocation proportional to the number of hospitals is preferred for several reasons:

- Fewer than 10 percent of government-planned database applications will produce nationwide estimates. The major government applications will investigate relationships among variables. For example, government researchers will do a substantial amount of regression modeling with these data.

- The HCUP-2 sample[2] used the same stratification and allocation scheme, and it has served AHCPR analysts well. Moreover, the large number of sample hospitals and discharges seemingly reduced the need for variance-reducing allocation schemes.

- AHCPR researchers wanted a simple, easily understood sampling methodology. It was an appealing idea that the NIS sample could be a "miniaturization" of the universe of hospitals (with the obvious geographical limitations imposed by data availability).

- AHCPR statisticians considered other optimal allocation schemes, including sampling hospitals with probabilities proportional to size (number of discharges), and they concluded that sampling with probability proportional to the number of hospitals was preferable. Even though it was recognized that the approach chosen would not be as efficient, the extremely large sample sizes yield good estimates. Furthermore, because the data are to be used for purposes other than producing national estimates, it is critical that all hospital types (including small hospitals) are adequately represented.

Overview of The Sampling Procedure

Once the universe of hospitals was stratified, up to 20 percent of the total number of U.S. hospitals was randomly selected within each stratum. If too few frame hospitals were in the stratum, then all frame hospitals were selected for the NIS, subject to sampling restrictions specified by states. To simplify variance calculations, at least two hospitals were drawn from each stratum. If fewer than two frame hospitals were contained in a stratum, then that stratum was merged with an "adjacent" stratum containing hospitals with similar characteristics.

A systematic random sample was drawn from each stratum, after sorting hospitals by state within each stratum, then by the three-digit zip code (the first three digits of the hospital's five-digit zip code) within each state, and then by a random number within each three-digit zip code. These sorts ensured further geographic generalizability of hospitals within the frame states, and random ordering of hospitals within three-digit zip codes.

Generally, three-digit zip codes that are near in value are geographically near within a state. Furthermore, the U.S. Postal Service locates regional mail distribution centers at the three-digit level. Thus, the boundaries tend to be a compromise between geographic size and population size.

1996 NIS Hospital Sampling Procedure

For the 1996 sample AHCPR decided that, rather than draw a new sample for 1996, the 1995 sample would be used to speed processing of the 1996 sample. Among the 938 hospitals selected for the 1995 NIS, 21 hospitals did not supply data for the 1996 HCUP. Among the remaining 917 hospitals, eight hospitals closed (according to the AHA) and three South Carolina hospitals had to be excluded because there was only one South Carolina hospital in three sampling frame strata. Therefore, 906 hospitals were included in the fifth release of the NIS. Details are shown in Table 41. The main consequence of this procedure is that hospitals new to the frame in 1996 could not enter the 1996 sample.

1995 NIS Hospital Sampling Procedure

The 1995 sample was drawn by a procedure that retained most of the 1994 hospitals, while allowing hospitals new to the frame an opportunity to enter the 1995 NIS.

Even in frame states that were present in the 1994 sample, hospitals that opened in 1995 needed a chance to enter the sample. Also, hospitals that changed strata between 1994 and 1995 were considered new to the 1995 frame.

Consequently, a recursive procedure was developed to update the sample from year to year in a way that properly accounted for changes in stratum size, composition, and sampling rate. The goal of this procedure was to maximize the year-to-year overlap among sample hospitals, yet keep the sampling rate constant for all hospitals *within a stratum*.

Table 41. Comparison of NIS Hospitals in 1995 and 1996

State	HOSPIDs in 1995 sample	NIS, Release 5 (1996)			
		HOSPIDs in 1995 and 1996 NIS	Excluded: No 1996 data	Excluded: Missing 1996 AHA info or restrictions	Percent of HOSPIDs excluded
Total	938	906	21	11	3.5%
AZ	15	15	0	0	0.0%
CA	105	103	2	0	1.9%
CO	22	21	1	0	4.8%
CT	9	8	0	1	12.5%
FL	141	138	1	2	2.2%
IA	54	53	1	0	1.9%
IL	73	72	0	1	1.4%
KS	61	60	1	0	1.7%
MA	25	19	6	0	31.6%
MD	39	39	0	0	0.0%
MO	49	47	2	0	4.3%
NJ	18	17	0	1	5.9%
NY	59	58	0	1	1.7%
OR	17	17	0	0	0.0%
PA	51	50	1	0	2.0%
SC	46	41	2	3	12.2%
TN	52	50	1	1	4.0%
WA	22	22	0	0	0.0%
WI	80	76	3	1	5.3%

The following procedure provides rules for creating a "year 2" sample, given that a "year 1" sample had already been drawn. In this example, year 1 would be 1994 and year 2 would be 1995. All notation is assumed to refer to sizes and probabilities within a particular stratum.

Probabilities P_1 and P_2 were calculated for sampling hospitals from the frame within the stratum for year 1 and year 2, respectively, based on the frame and universe for year 1 and year 2, respectively. These probabilities were set by the same algorithm used to calculate P for the 1988 hospital sample (see Technical Supplement: *Design of the HCUP Nationwide Inpatient Sample, Release 1*, section "1988 NIS Hospital Sampling Procedure.")

Now consider the three possibilities associated with changes between years 1 and 2 in the stratum-specific hospital sampling probabilities:

1. $P_2 = P_1$: The target probability was unchanged.

2. $P_2 < P_1$: The target probability decreased.

3. $P_2 > P_1$: The target probability increased.

Below is the procedure used for each of these three cases with one exception: if the stratum-specific probability of selection P_2 was equal to 1, then all frame hospitals were selected for the year 2 sample, regardless of the value of P_1.

Stratum-Specific Sampling Rates the Same ($P_2 = P_1$). If the probability P_2 was the same as P_1, all hospitals in the year 1 sample that remained in the year 2 frame were retained for the year 2 sample. Any new frame hospitals (those in the year 2 frame but not in the year 1 frame) were selected at the rate P_2, using the systematic sampling method described for the 1988 sample selection in Technical Supplement: *Design of the HCUP Nationwide Inpatient Sample, Release 1*.

Stratum-Specific Sampling Rate Decreased ($P_2 < P_1$). Now consider the case where the probability of selection decreased between years 1 and 2. First, hospitals new to the frame were sampled with probability P_2. Second, hospitals previously selected for the year 1 sample (that remained in the year 2 frame) were selected for the year 2 sample with probability $P_2 \div P_1$.

The justification for this second procedure was straightforward. For the year 1 sample hospitals that stayed in the frame, the year 1 sample was viewed as the first stage of a two-stage sampling process. The first stage was carried out at the sampling rate of P_1. The second stage was carried out at the sampling rate of $P_2 \div P_1$. Consequently, the "overall" probability of selection was $P_1 \text{ x } P_2 \div P_1 = P_2$.

Stratum-Specific Sampling Rate Increased ($P_2 > P_1$). The procedures associated with the case in which the probability of selection was increased between year 1 and year 2 were equally straightforward. First, hospitals new to the frame were sampled with probability P_2. Second, hospitals that were selected in year 1 (that remained in the year 2 frame) were selected for the year 2 sample. Third, hospitals that were in the frame for both years 1 and 2, but not selected for the year 1 sample, were selected for the year 2 sample with probability $(P_2\text{-}P_1) \div (1\text{-}P_1)$.

The justification for this sampling rate, $(P_2\text{-}P_1) \div (1\text{-}P_1)$, is somewhat complex. In year 1 certain frame hospitals were included in the sample at the rate P_1. This can also be viewed as having excluded a set of hospitals at the rate $(1\text{-}P_1)$. Likewise, in year 2 it was imperative that each hospital excluded from the year 1 sample be excluded from the year 2 sample at an overall rate of $(1\text{-}P_2)$.

Since $P_2 > P_1$, then $(1\text{-}P_2) < (1\text{-}P_1)$. Therefore, just as was done for the case of $P_2 < P_1$, multistage selection was implemented. However, it was implemented for exclusion rather than inclusion.

Therefore, those hospitals excluded from the year 1 sample were also excluded from the year 2 sample at the rate $S = (1-P_2) \div (1-P_1)$. This gave them the desired overall *exclusion* rate of $(1-P_1) \times (1-P_2) \div (1-P_1) = (1-P_2)$. Consequently, the *inclusion* rate for these hospitals was set at $1-S = (P_2-P_1) \div (1-P_1)$.

Zero-Weight Hospitals

Beginning in 1993, the NIS samples contain no zero-weight hospitals. For a description of zero-weight hospitals in the 1988-1992 sample, see the Technical Supplement: *Design of the HCUP Nationwide Inpatient Sample, Release 1*.

Ten Percent Subsamples

Two nonoverlapping 10 percent subsamples of discharges were drawn from the NIS file for each year. The subsamples were selected by drawing every tenth discharge starting with two different starting points (randomly selected between 1 and 10). Having a different starting point for each of the two subsamples guaranteed that they would not overlap. Discharges were sampled so that 10 percent of each hospital's discharges in each quarter were selected for each of the subsamples. The two samples can be combined to form a single, generalizable 20 percent subsample of discharges.

FINAL HOSPITAL SAMPLE

The annual numbers of hospitals and discharges in each release of the NIS are shown in Table 42 for both the regular NIS sample and the total sample (which includes zero-weight hospitals for 1988-1992).

Table 42. NIS Hospital Sample

Year	Regular Sample		Total Sample	
	Number of Hospitals	**Number of Discharges**	**Number of Hospitals**	**Number of Discharges**
NIS, Release 1				
1988	758	5,242,904	759	5,265,756
1989	875	6,067,667	882	6,110,064
1990	861	6,156,638	871	6,268,515
1991	847	5,984,270	859	6,156,188
1992	838	6,008,001	856	6,195,744
NIS, Release 2				
1993	913	6,538,976	913	6,538,976
NIS, Release 3				
1994	904	6,385,011	904	6,385,011
NIS, Release 4				
1995	938	6,714,935	938	6,714,935
NIS, Release 5				
1996	906	6,542,069	906	6,542,069
Total		55,640,471		56,177,258

A more detailed breakdown of the 1996 NIS hospital sample by geographic region is shown in Table 43. For each geographic region, Table 43 shows the number of:

- universe hospitals (Universe),
- frame hospitals (Frame),
- sampled hospitals (Sample),
- target hospitals (Target = 20 percent of the universe), and
- shortfall hospitals (Shortfall = Sample - Target).

Table 43. Number of Hospitals in Universe, Frame, Regular Sample, Target, and Shortfall By Region, 1996

Region	Universe	Frame	Sample	Target	Shortfall
NE	753	631	152	151	1
MW	1488	547	308	298	10
S	1,975	391	268	395	-127
W	966	699	178	193	-15
Total	5,182	2,268	906	1037	-131

For example, in 1996 the Northeast region contained 753 hospitals in the universe. It also contained 631 hospitals in the frame, of which 152 hospitals were drawn for the sample. This was 1 hospital more than the target sample size of 151.

Table 44 shows the number of hospitals in the universe, frame, and regular sample for each state in the sampling frame for 1996. In all states except Illinois, South Carolina, Tennessee and Missouri, the difference between the universe and the frame represents the difference in the number of community hospitals in the 1996 AHA Annual Survey of Hospitals and the number of community hospitals for which data were supplied to HCUP. As explained earlier, the number of hospitals in the Illinois frame is approximately 53 percent of the hospitals in the Illinois universe in order to comply with the agreement with the data source concerning the restriction on the number of Illinois discharges. The number of hospitals in the South Carolina frame is eleven fewer than the South Carolina universe. Six hospitals were excluded because of sampling restrictions stipulated by South Carolina, and five hospitals were not included in the data supplied to HCUP. The number of hospitals in the Tennessee frame is 38 fewer than the Tennessee universe. Four hospitals were excluded because of sampling restrictions stipulated by Tennessee, and 34 hospitals were not included in the data supplied to HCUP. The number of hospitals in the Missouri frame is 48 fewer than the Missouri universe. Thirty-five hospitals were excluded because they signed release for confidential use only, and 13 hospitals were not included in the data supplied to HCUP.

The number of hospitals in the NIS hospital samples that continue across multiple sample years is shown in Table 44. This table will be of interest to those who may combine Releases 1, 2, 3, 4 and 5 of the NIS. Table 45 shows that longitudinal cohorts that span several years and include 1988 and 1993 are the lowest in number of continuing sample hospitals. For example, if 1988 is taken as a starting year, only 37.3 percent of the 1988 hospital sample continued in the 1996 sample (283 of 758).

Table 44. Number of Hospitals in the Universe, Frame, and Sample for States in the Sampling Frame: 1996

State	Universe	Frame	Sample
AZ	62	62	15
CA	421	418	103
CO	68	67	21
CT	33	31	8
FL	213	197	138
IA	115	115	53
IL	205	108	72
KS	132	121	60
MA	88	75	19
MD	51	51	39
MO	127	79	47
NJ	89	82	17
NY	227	226	58
OR	63	62	17
PA	223	217	50
SC	66	55	41
TN	126	88	50
WA	91	90	22
WI	124	124	76
Total	2524	2268	906

Table 45. Number of Hospitals and Discharges in Longitudinal Cohort

Number of Years	Calendar Years	Longitudinal Regular Sample Hospitals	% of Base Year Sample	Longitudinal Regular Sample Discharges
2	1988-1989	610	80.5	8,492,039
	1989-1990	815	93.1	11,525,749
	1990-1991	802	93.1	11,297,175
	1991-1992	781	92.2	11,272,981
	1992-1993	609	72.7	8,804,638
	1993-1994	693	75.9	10,271,404
	1994-1995	762	84.3	10,747,682
3	1988-1990	573	75.6	12,168,677
	1989-1991	763	87.2	16,074,381
	1990-1992	745	86.5	16,085,651
	1991-1993	570	67.3	12,559,421
	1992-1994	540	64.4	11,279,667
	1993-1995	598	65.5	13,241,070
	1994-1996	740	81.9	15,651,230
4	1988-1991	542	71.5	15,096,807
	1989-1992	709	81.0	20,340,970
	1990-1993	548	63.6	16,023,500
	1991-1994	508	60.0	14,481,319
	1992-1995	464	55.4	12,712,613
5	1988-1992	502	66.2	18,106,098
	1989-1993	523	59.8	19,000,777
	1990-1994	490	56.9	17,437,229
	1991-1995	439	51.8	15,405,253
6	1988-1993	378	49.9	16,906,818
	1989-1994	471	53.8	19,987,910
	1990-1995	422	49.0	14,817,797
7	1988-1994	335	44.2	17,128,064
	1989-1995	408	46.6	19,924,107
8	1988-1995	289	38.1	16,658,485
9	1988-1996	283	37.3	18,576,353

SAMPLING WEIGHTS

Although the sampling design was simple and straightforward, it is necessary to incorporate sample weights to obtain state and national estimates. Therefore, sample weights were developed separately for hospital- and discharge-level analyses. Three hospital-level weights were developed to weight NIS sample hospitals to the state, frame, and universe. Similarly, three discharge-level weights were developed to weight NIS sample discharges to the state, frame, and universe.

Hospital-Level Sampling Weights

Universe Hospital Weights. Hospital weights to the universe were calculated by post-stratification. For each year, hospitals were stratified on the same variables that were used for sampling: geographic region, urban/rural location, teaching status, bedsize, and control. The strata that were collapsed for sampling were also collapsed for sample weight calculations. Within stratum s, each NIS sample hospital's universe weight was calculated as:

$$W_s(\text{universe}) = N_s(\text{universe}) \div N_s(\text{sample}),$$

where N_s(universe) and N_s(sample) were the number of community hospitals within stratum s in the universe and sample, respectively. Thus, each hospital's universe weight is equal to the number of universe hospitals it represented during that year.

Frame Hospital Weights. Hospital-level sampling weights were also calculated to represent the entire collection of states in the frame using the same post-stratification scheme as described above for the weights to represent the universe. For each year, within stratum s, each NIS sample hospital's frame weight was calculated as:

$$W_s(\text{frame}) = N_s(\text{frame}) \div N_s(\text{sample}).$$

N_s(frame) was the total number of universe community hospitals within stratum s in the states that contributed data to the frame. N_s(sample) was the number of sample hospitals selected for the NIS in stratum s. Thus, each hospital's frame weight is equal to the number of universe hospitals it represented in the frame states during that year.

State Hospital Weights. For each year, a hospital's weight to its state was calculated in a similar fashion. Within each state, strata often had to be collapsed after sample selection for development of weights to ensure a minimum of two sample hospitals within each stratum. For each state and each year, within stratum s, each NIS sample hospital's state weight was calculated as:

$$W_s(\text{state}) = N_s(\text{state}) \div N_s(\text{state sample}).$$

N_s(state) was the number of universe community hospitals in the state within stratum s. N_s(state sample) was the number of hospitals selected for the NIS from that state in stratum s. Thus, each hospital's state weight is equal to the number of hospitals that it represented in its state during that year.

All of these hospital weights can be rescaled if necessary for selected analyses, to sum to the NIS hospital sample size each year.

Discharge-Level Sampling Weights

The calculations for discharge-level sampling weights were very similar to the calculations of hospital-level sampling weights. The discharge weights usually are constant for all discharges within a stratum.

The only exceptions were for strata with sample hospitals that, according to the AHA files, were open for the entire year but contributed less than their full year of data to the NIS. For those hospitals, we *adjusted* the number of observed discharges by a factor 4 ÷ Q, where Q was the number of calendar quarters that the hospital contributed discharges to the NIS. For example, when a sample hospital contributed only two quarters of discharge data to the NIS, the *adjusted* number of discharges was double the observed number.

With that minor adjustment, each discharge weight is essentially equal to the number of reference (universe, frame, or state) discharges that each sampled discharge represented in its stratum. This calculation was possible because the number of total discharges was available for every hospital in the universe from the AHA files. Each universe hospital's AHA discharge total was calculated as the sum of newborns and total facility discharges.

Universe Discharge Weights. Discharge weights to the universe were calculated by post-stratification. Hospitals were stratified just as they were for universe hospital weight calculations. Within stratum s, for hospital i, each NIS sample discharge's universe weight was calculated as:

$$DW_{is}(\text{universe}) = [DN_s(\text{universe}) \div ADN_s(\text{sample})] * (4 \div Q_i),$$

where DN_s(universe) was the number of discharges from community hospitals in the universe within stratum s; ADN_s(sample) was the number of *adjusted* discharges from sample hospitals selected for the NIS; and Q_i was the number of quarters of discharge data contributed by hospital i to the NIS (usually $Q_i = 4$). Thus, each discharge's weight is equal to the number of universe discharges it represented in stratum s during that year.

Frame Discharge Weights. Discharge-level sampling weights were also calculated to represent all discharges from the entire collection of states in the frame using the same post-stratification scheme described above for the discharge weights to represent the universe. For each year, within stratum s, for hospital i, each NIS sample discharge's frame weight was calculated as:

$$W_{is}(\text{frame}) = [DN_s(\text{frame}) \div ADN_s(\text{sample})] * (4 \div Q_i),$$

DN_s(frame) was the number of discharges from all community hospitals in the states that contributed to the frame within stratum s. ADN_s(sample) was the number of *adjusted* discharges from sample hospitals selected for the NIS in stratum s. Q_i was the number of quarters of discharge data contributed by hospital i to the NIS (usually $Q_i = 4$). Thus, each discharge's frame weight is equal to the number of discharges it represented in the frame states during that year.

State Discharge Weights. A discharge's weight to its state was similarly calculated. Strata were collapsed in the same way as they were for the state hospital weights to ensure a minimum of two sample hospitals within each stratum. Within stratum s, for hospital i, each NIS sample discharge's state weight was calculated as:

$$W_{is}(\text{state}) = [DN_s(\text{state}) \div ADN_s(\text{state sample})] * (4 \div Q_i),$$

DN_s(state) was the number of discharges from all community hospitals in the state within stratum s. ADN_s(state sample) was the *adjusted* number of discharges from hospitals selected for the NIS from that state in stratum s. Q_i was the number of quarters of discharge data contributed by hospital i to the NIS (usually $Q_i = 4$). Thus, each discharge's state weight is equal to the number of discharges that it represented in its state during that year.

All of these discharge weights can be rescaled if necessary for selected analyses, to sum to the NIS discharge sample size each year.

Discharge Weights for 10 Percent Subsamples

In the 10 percent subsamples, each discharge had a 10 percent chance of being drawn. Therefore, the discharge weights contained in the Hospital Weights file can be multiplied by 10 for each of the subsamples, or multiplied by 5 for the two subsamples combined.

DATA ANALYSIS

Variance Calculations

It may be important for researchers to calculate a measure of precision for some estimates based on the NIS sample data. Variance estimates must take into account both the sampling design and the form of the statistic. The sampling design was a stratified, single-stage cluster sample. A stratified random sample of hospitals (clusters) were drawn and then *all* discharges were included from each selected hospital.

If hospitals inside the frame were similar to hospitals outside the frame, the sample hospitals can be treated as if they were randomly selected from the entire universe of hospitals within each stratum. Standard formulas for a stratified, single-stage cluster sampling without replacement could be used to calculate statistics and their variances in most applications.

A multitude of statistics can be estimated from the NIS data. Several computer programs are listed below that calculate statistics and their variances from sample survey data. Some of these programs use general methods of variance calculations (e.g., the jackknife and balanced half-sample replications) that take into account the sampling design. However, it may be desirable to calculate variances using formulas specifically developed for some statistics.

In most cases, computer programs are readily available to perform these calculations. For instance, OSIRIS IV, developed at the University of Michigan, and SUDAAN, developed at the Research Triangle Institute, do calculations for numerous statistics arising from the stratified, single-stage cluster sampling design. An example of using SUDAAN to calculate variances in the NIS is presented in Technical Supplement: *Calculating Variances Using Data from the HCUP Nationwide Inpatient Sample*.[3]

These variance calculations are based on finite-sample theory, which is an appropriate method for obtaining cross-sectional, nationwide estimates of outcomes. According to finite-sample theory, the intent of the estimation process is to obtain estimates that are precise representations of the nationwide population at a specific point in time. In the context of the NIS, any estimates that attempt to accurately describe characteristics (such as expenditure and utilization patterns or hospital market factors) and interrelationships among characteristics of hospitals and discharges during a specific year from 1988 to 1996 should be governed by finite-sample theory.

Alternatively, in the study of hypothetical population outcomes not limited to a specific point in time, analysts may be less interested in specific characteristics from the finite population (and time period) from which the *sample* was drawn, than they are in hypothetical characteristics of a conceptual "superpopulation" from which any particular finite *population* in a given year might have been drawn. According to this superpopulation model, the nationwide population in a given year is only a snapshot in time of the possible interrelationships among hospital, market, and discharge characteristics. In a given year, all possible interactions between such characteristics may not have been observed, but analysts may wish to predict or simulate interrelationships that may occur in the future.

Under the finite-population model, the variances of estimates approach zero as the sampling fraction approaches one, since the population is defined at that point in time, and because the estimate is for a characteristic as it existed at the time of sampling. This is in contrast to the superpopulation model, which adopts a stochastic viewpoint rather than a deterministic viewpoint. That is, the nationwide population in a particular year is viewed as a random sample of some underlying superpopulation over time.

Different methods are used for calculating variances under the two sample theories. Under the superpopulation (stochastic) model, procedures (such as those described by Potthoff, Woodbury, and Manton[4]) have been developed to draw inferences using weights from complex samples. In this context, the survey weights are not used to weight the sampled cases to the universe, because the universe is conceptually infinite in size. Instead, these weights are used to produce unbiased estimates of parameters that govern the superpopulation.

In summary, the choice of an appropriate method for calculating variances for nationwide estimates depends on the type of measure and the intent of the estimation process.

Computer Software for Variance Calculations

The hospital weights will be useful for producing hospital-level statistics for analyses that use the *hospital* as the unit of analysis, and the discharge weights will be useful for producing discharge-level statistics for analyses that use the *discharge* as the unit of analysis. These would be used to weight the sample data in estimating population statistics.

Several statistical programming packages allow weighted analyses.[5] For example, nearly all SAS (Statistical Analysis System) procedures incorporate weights.

In addition, several publicly available subroutines have been developed specifically for calculating statistics and their standard errors from survey data:

- OSIRIS IV was developed by L. Kish, N. Van Eck, and M. Frankel at the Survey Research Center, University of Michigan. It consists of two main programs for estimating variances from complex survey designs.

- SUDAAN, a set of SAS subroutines, was developed at the Research Triangle Institute by B. V. Shah. It is adequate for handling most survey designs with stratification. The procedures can handle estimation and variance estimation for means, proportions, ratios, and regression coefficients.

- SUPER CARP (Cluster Analysis and Regression Program) was developed at Iowa State University by W. Fuller, M. Hidiroglou, and R. Hickman. This program computes estimates and variance estimates for multistage, stratified sampling designs with arbitrary probabilities of selection. It can handle estimated totals, means, ratios, and regression estimates.

The NIS database includes a Hospital Weights file with variables required by these programs to calculate finite population statistics. In addition to the sample weights described earlier, hospital identifiers (PSUs), stratification variables, and stratum-specific totals for the numbers of discharges and hospitals are included so that finite-population corrections (FPCs) can be applied to variance estimates.

In addition to these subroutines, standard errors can be estimated by validation and cross-validation techniques. Given that a very large number of observations will be available for most analyses, it may be feasible to set aside a part of the data for validation purposes. Standard errors and confidence intervals can then be calculated from the validation data. If the analytical file is too small to set aside a large validation sample, cross-validation techniques may be used.

For example, tenfold cross-validation would split the data into ten equal-sized subsets. The estimation would take place in ten iterations. At each iteration, the outcome of interest is predicted for one-tenth of the observations by an estimate based on a model fit to the other nine-tenths of the observations. Unbiased estimates of error variance are then obtained by comparing the actual values to the predicted values obtained in this manner.

Finally, it should be noted that a large array of hospital-level variables are available for the entire universe of hospitals, including those outside the sampling frame. For instance, the variables from the AHA surveys and from the Medicare Cost Reports are available for nearly all hospitals. To the extent that hospital-level outcomes correlate with these variables, they may be used to sharpen regional and nationwide estimates.

As a simple example, each hospital's number of C-sections would be correlated with their total number of deliveries. The number of C-sections must be obtained from discharge data, but the number of deliveries is available from AHA data. Thus, if a regression can be fit predicting C-sections from deliveries based on the NIS data, that regression can then be used to obtain hospital-specific estimates of the number of C-sections for all hospitals in the universe.

Longitudinal Analyses

As previously shown in Table 45, hospitals that continue in the NIS for multiple consecutive years are a subset of the hospitals in the NIS for any one of those years. Consequently, longitudinal analyses of hospital-level outcomes may be biased if they are based on any subset of NIS hospitals limited to continuous NIS membership. In particular, such subsets would tend to contain fewer hospitals that opened, closed, split, merged, or changed strata. Further, the sample weights were developed as annual, cross-sectional weights rather than longitudinal weights. Therefore, different weights might be required, depending on the statistical methods employed by the analyst.

One approach to consider in hospital-level longitudinal analyses is to use repeated-measure models that allow hospitals to have missing values for some years. However, the data are not actually missing for some hospitals, such as those that closed during the study period. In any case, the analyses may be more efficient (e.g., produce more precise estimates) if they account for the potential correlation between repeated measures on the same hospital over time, yet incorporate data from all hospitals in the sample during the study period.

Discharge Subsamples

The two nonoverlapping 10 percent subsamples of discharges were drawn from the NIS file for each year for several reasons pertaining to data analysis. One reason for creating the subsamples was to reduce processing costs for selected studies that will not require the entire NIS. Another reason is that the two subsamples may be used to validate models and obtain unbiased estimates of standard errors. That is, one subsample may be used to estimate statistical models, and the other subsample may be used to test the fit of those models on new data. This is a very important analytical step, particularly in exploratory studies, where one runs the risk of fitting noise.

For example, it is well known that the percentage of variance explained by a regression, R^2, is generally overestimated by the data used to fit a model. The regression model could be estimated from the first subsample and then applied to the second subsample. The squared correlation between the actual and predicted value in the second subsample is an unbiased estimate of the model's true explanatory power when applied to new data.

ENDNOTES

1. Most AHA surveys do not cover a January-to-December calendar year. The number of hospitals for 1988-1991 are based on the HCUP calendar-year version of the AHA Annual Survey files. To create a calendar-year reporting period, data from the AHA surveys must be apportioned in some manner across calendar years. Survey responses were converted to calendar-year periods for 1988-1991 by merging data from adjacent survey years. The number of hospitals for 1992-1994 are based on the AHA Annual Survey files.

2. Coffey, R. and D. Farley (1988, July). *HCUP-2 Project Overview,* (DHHS Publication No. (PHS) 88-3428. Hospital Studies Program Research Note 10, National Center for Health Services Research and Health Care Technology Assessment, Rockville, MD: Public Health Service.

3. Duffy, S.Q. and J.P. Sommers (1996, March). *Calculating Variances Using Data from the HCUP Nationwide Inpatient Sample*. Rockville, MD: Agency for Health Care Policy and Research.

4. Potthoff, R.F., M.A. Woodbury, and K.G. Manton (1992). "Equivalent Sample Size" and "Equivalent Degrees of Freedom" Refinements for Inference Using Survey Weights Under Superpopulation Models. *Journal of the American Statistical Association*, Vol. 87, 383-396.

5. Carlson, B.L., A.E. Johnson, and S.B. Cohen (1993). An Evaluation of the Use of Personal Computers for Variance Estimation with Complex Survey Data. *Journal of Official Statistics*, Vol. 9, No. 4, 795-814.

TECHNICAL SUPPLEMENT 10:
DESIGN OF THE HCUP NATIONWIDE INPATIENT SAMPLE, RELEASE 6

INTRODUCTION

The Nationwide Inpatient Sample (NIS) of the Healthcare Cost and Utilization Project (HCUP) was established to provide analyses of hospital utilization across the United States. The target universe includes all acute-care discharges from all community hospitals in the United States; the NIS comprises all discharges from a sample of hospitals in this target universe.

NIS Release	Calendar Year	States	Sample Hospitals	Sample Discharge (millions)
1	1988–1992	8–11	758–875	5.2–6.2
2	1993	17	913	6.5
3	1994	17	904	6.4
4	1995	19	938	6.7
5	1996	19	906	6.5
6	1997	22	1012	7.1

Thus, the NIS supports both cross-sectional and longitudinal analyses.

Potential research issues focus on both discharge- and hospital-level outcomes. Discharge outcomes of interest include trends in inpatient treatments with respect to:

- frequency,
- costs,
- lengths of stay,
- effectiveness,
- quality of care,
- appropriateness, and
- access to hospital care.

Hospital outcomes of interest include:

- mortality rates,
- complication rates,
- patterns of care,
- diffusion of technology, and
- trends toward specialization.

These and other outcomes are of interest for the nation as a whole and for policy-relevant inpatient subgroups defined by geographic regions, patient demographics, hospital characteristics, physician characteristics, and pay sources.

This report provides a detailed description of the NIS, Release 6 sample design, as well as a summary of the resultant hospital sample. Sample weights were developed to obtain national estimates of hospital and inpatient parameters. These weights and other special-use weights are described in detail. Tables include cumulative information for all six NIS releases to provide a longitudinal view of the database.

THE NIS HOSPITAL UNIVERSE

The hospital universe is defined by all hospitals that were open during any part of the calendar year and were designated as community hospitals in the American Hospital Association (AHA) Annual Survey of Hospitals. For purposes of the NIS, the definition of a community hospital is that used by the AHA: "all nonfederal short-term general and other specialty hospitals, excluding hospital units of institutions." Consequently, Veterans Hospitals and other federal hospitals are excluded. Table 46 shows the number of universe hospitals for each year based on the AHA Annual Survey.

Table 46. Hospital Universe[1]

Year	Number of Hospitals
1988	5,607
1989	5,548
1990	5,468
1991	5,412
1992	5,334
1993	5,313
1994	5,290
1995	5,260
1996	5,182
1997	5,113

Hospital Merges, Splits, and Closures

All hospital entities that were designated community hospitals in the AHA hospital file were included in the hospital universe. Therefore, if two or more community hospitals merged to create a new community hospital, the original hospitals and the newly-formed hospital were all considered separate hospital entities in the universe for the year of the merge. Likewise, if a community hospital split, the original hospital and all newly created community hospitals were separate entities in the universe for the year of the split. Finally, community hospitals that closed during a year were included as long as they were in operation during some part of the calendar year.

Stratification Variables

To help ensure representativeness, sampling strata were defined based on five hospital characteristics contained in the AHA hospital files. The stratification variables were as follows:

1) *Geographic Region – Northeast, Midwest, West, and South*. This is an important stratifier because practice patterns have been shown to vary substantially by region. For example, lengths of stay tend to be longer in East Coast hospitals than in West Coast hospitals.

2) *Control – government nonfederal, private not-for-profit, and private investor-owned.* These types of hospitals tend to have different missions and different responses to government regulations and policies.

3) *Location – urban or rural.* Government payment policies often differ according to this designation. Also, rural hospitals are generally smaller and offer fewer services than urban hospitals.

4) *Teaching Status – teaching or nonteaching*. The missions of teaching hospitals differ from nonteaching hospitals. In addition, financial considerations differ between these two hospital groups. Currently, the Medicare DRG payments are uniformly higher to teaching hospitals than to nonteaching hospitals. A hospital is considered to be a teaching hospital if it has an AMA-approved residency program or is a member of the Council of Teaching Hospitals (COTH).

5) *Bedsize – small, medium, and large.* Bedsize categories are based on hospital beds, and are specific to the hospital's location and teaching status, as shown in Table 47.

Table 47. Bedsize Categories

Location and Teaching Status	Hospital Bedsize		
	Small	**Medium**	**Large**
Rural	1-49	50-99	100+
Urban, nonteaching	1-99	100-199	200+
Urban, teaching	1-299	300-499	500+

Rural hospitals were not split according to teaching status, because rural teaching hospitals were rare. For example, in 1988 there were only 20 rural teaching hospitals. The bedsize categories were defined within location and teaching status because they would otherwise have been redundant. Rural hospitals tend to be small; urban nonteaching hospitals tend to be medium-sized; and urban teaching hospitals tend to be large. Yet it was important to recognize gradations of size within these types of hospitals.

For example, in serving rural discharges, the role of "large" rural hospitals (particularly rural referral centers) often differs from the role of "small" rural hospitals. The cut-off points for the bedsize categories are consistent with those used in *Hospital Statistics,* published annually by the AHA.

To further ensure geographic representativeness, implicit stratification variables included state and three-digit zip code (the first three digits of the hospital's five-digit zip code). The hospitals were sorted according to these variables prior to systematic sampling.

HOSPITAL SAMPLING FRAME

For each year, the *universe* of hospitals was established as all community hospitals located in the U.S. However, it was not feasible to obtain and process all-payer discharge data from a random sample of the entire universe of hospitals for at least two reasons. First, all-payer discharge data were not available from all hospitals for research purposes. Second, based on the experience of prior hospital discharge data collections, it would have been too costly to obtain data from individual hospitals, and it would have been too burdensome to process each hospital's unique data structure.

Therefore, the NIS *sampling frame* was constructed from the subset of universe hospitals that released their discharge data for research use. Two sources for all-payer discharge data were state agencies and private data organizations, primarily state hospital associations. At the time when the sample was drawn, the Agency for Health Care Policy and Research (AHCPR) had agreements with 22 data sources that maintain statewide, all-payer discharge data files to include their data in the HCUP database. However, only 8 states in 1988 and 11 states in 1989-1992 could be included in the first release of the NIS, an additional 6 states were included in the second and the third release of the NIS, another 2 states were included in the fourth and the fifth releases of the NIS, and 3 more states were included in this sixth release of the NIS as shown in Table 48.

Table 48. States in the Frame for NIS Releases

Years	States in the Frame
NIS, Release 1	
1988	California, Colorado, Florida, Iowa, Illinois, Massachusetts, New Jersey, and Washington
1989-1992	Add Arizona, Pennsylvania, and Wisconsin
NIS, Release 2	
1993	Add Connecticut, Kansas, Maryland, New York, Oregon, South Carolina
NIS, Release 3	
1994	No new additions
NIS, Release 4	
1995	Add Missouri, Tennessee
NIS, Release 5	
1996	No new additions
NIS, Release 6	
1997	Add Georgia, Hawaii, and Utah

The list of the entire frame of hospitals was composed of all AHA community hospitals in each of the frame states *that could be matched to the discharge data provided to HCUP*. If an AHA community hospital could not be matched to the discharge data provided by the data source, it was eliminated from the sampling frame (but not from the target universe). Further restrictions were put on the sampling frames for Georgia, Hawaii, Illinois, South Carolina, Missouri, and Tennessee.

The Illinois Health Care Cost Containment Council stipulated that no more than 40 percent of the discharges provided by Illinois could be included in the database for any calendar quarter. Consequently, a systematic random sample of Illinois hospitals was drawn for the 1997 frame. This prevented the sample from including more than 40 percent of Illinois discharges.

Georgia, Hawaii, South Carolina and Tennessee stipulated that only hospitals that appear in sampling strata with two or more hospitals were to be included in the NIS.

Due to this restriction, one Georgia hospital, six Hawaii hospitals, six South Carolina hospitals and five Tennessee hospitals were excluded from the 1997 frame leaving 158 Georgia community hospitals, 11 Hawaii hospitals, 54 South Carolina hospitals and 92 Tennessee community hospitals in the 1997 frame.

Missouri stipulated that only hospitals that had signed releases for public use should be included in the NIS. For 1997, thirty-five Missouri hospitals signed releases for confidential use only. These hospitals were excluded from the sampling frame, leaving 75 hospitals in the 1997 frame.

The number of frame hospitals for each year is shown in Table 49.

Table 49. Hospital Frame

Year	Number of Hospitals
1988	1,247
1989	1,658
1990	1,620
1991	1,604
1992	1,591
1993	2,168
1994	2,135
1995	2,284
1996	2,268
1997	2,452

HOSPITAL SAMPLE DESIGN

Design Requirements

The NIS is a stratified probability sample of hospitals in the frame, with sampling probabilities calculated to select 20 percent of the universe contained in each stratum. The overall objective was to select a sample of hospitals "generalizable" to the target universe, which includes hospitals outside the frame (zero probability of selection). Moreover, this sample was to be geographically dispersed, yet drawn from the subset of states with inpatient discharge data that agreed to provide such data to the project.

It should be possible, for example, to estimate DRG-specific average lengths of stay over all U.S. hospitals using weighted average lengths of stay, based on averages or regression estimates from the NIS. Ideally, relationships among outcomes and their correlates estimated from the NIS should generally hold across all U.S. hospitals. However, since only 22 states contributed data to this sixth release, some estimates may differ from estimates from comparative data sources. When possible, estimates based on the NIS should be checked against national benchmarks, such as Medicare data or data from the National Hospital Discharge Survey to determine the appropriateness of the NIS for specific analyses (see the Technical Supplement: *Comparative Analysis of HCUP and NHDS Inpatient Discharge Data*).

The target sample size was 20 percent of the total number of community hospitals in the U.S. for 1997. This sample size was determined by AHCPR based on their experience with similar research databases.

Alternative stratified sampling allocation schemes were considered. However, allocation proportional to the number of hospitals is preferred for several reasons:

- Fewer than 10 percent of government-planned database applications will produce nationwide estimates. The major government applications will investigate relationships among variables. For example, government researchers will do a substantial amount of regression modeling with these data.

- The HCUP-2 sample[2] used the same stratification and allocation scheme, and it has served AHCPR analysts well. Moreover, the large number of sample hospitals and discharges seemingly reduced the need for variance-reducing allocation schemes.

- AHCPR researchers wanted a simple, easily understood sampling methodology. It was an appealing idea that the NIS sample could be a "miniaturization" of the universe of hospitals (with the obvious geographical limitations imposed by data availability).

- AHCPR statisticians considered other optimal allocation schemes, including sampling hospitals with probabilities proportional to size (number of discharges), and they concluded that sampling with probability proportional to the number of hospitals was preferable. Even though it was recognized that the approach chosen would not be as efficient, the extremely large sample sizes yield good estimates. Furthermore, because the data are to be used for purposes other than producing national estimates, it is critical that all hospital types (including small hospitals) are adequately represented.

Overview of the Sampling Procedure

Once the universe of hospitals was stratified, up to 20 percent of the total number of U.S. hospitals was randomly selected within each stratum. If too few frame hospitals were in the stratum, then all frame hospitals were selected for the NIS, subject to sampling restrictions specified by states. To simplify variance calculations, at least two hospitals were drawn from each stratum. If fewer than two frame hospitals were contained in a stratum, then that stratum was merged with an "adjacent" stratum containing hospitals with similar characteristics.

A systematic random sample was drawn from each stratum, after sorting hospitals by state within each stratum, then by the three-digit zip code (the first three digits of the hospital's five-digit zip code) within each state, and then by a random number within each three-digit zip code. These sorts ensured further geographic generalizability of hospitals within the frame states, and random ordering of hospitals within three-digit zip codes.

Generally, three-digit zip codes that are near in value are geographically near within a state. Furthermore, the U.S. Postal Service locates regional mail distribution centers at the three-digit level. Thus, the boundaries tend to be a compromise between geographic size and population size.

1997 NIS Hospital Sampling Procedure

The 1997 sample was drawn by a procedure that retained most of the 1995 hospitals while allowing hospitals new to the frame an opportunity to enter the 1997 NIS. (Note: The 1996 NIS was not selected through a sampling procedure but was constructed as the subset of 1995 NIS hospitals that continued to supply data in 1996).

Even in frame states that were present in the 1995 sample, hospitals that opened in 1997 needed a chance to enter the sample. Also, hospitals that changed strata between 1995 and 1997 were considered new to the 1997 frame.

Consequently, a recursive procedure was developed to update the sample from year to year in a way that properly accounted for changes in stratum size, composition, and sampling rate. The goal of this procedure was to maximize the year-to-year overlap among sample hospitals, yet keep the sampling rate constant for all hospitals *within a stratum*.

The following procedure provides rules for creating a "year 2" sample, given that a "year 1" sample had already been drawn. In this example, year 1 would be 1995 and year 2 would be 1997. All notation is assumed to refer to sizes and probabilities within a particular stratum.

Probabilities P_1 and P_2 were calculated for sampling hospitals from the frame within the stratum for year 1 and year 2, respectively, based on the frame and universe for year 1 and year 2, respectively. These probabilities were set by the same algorithm used to calculate P for the original (1988) hospital sample (see Technical Supplement: *Design of the HCUP Nationwide Inpatient Sample, Release 1*, section "1988 NIS Hospital Sampling Procedure.")

Now consider the three possibilities associated with changes between years 1 and 2 in the stratum-specific hospital sampling probabilities:

1. $P_2 = P_1$: The target probability was unchanged.

2. $P_2 < P_1$: The target probability decreased.

3. $P_2 > P_1$: The target probability increased.

Below is the procedure used for each of these three cases with one exception: if the stratum-specific probability of selection P_2 was equal to 1, then all frame hospitals were selected for the year 2 sample, regardless of the value of P_1.

Stratum-Specific Sampling Rates the Same ($P_2 = P_1$). If the probability P_2 was the same as P_1, all hospitals in the year 1 sample that remained in the year 2 frame were retained for the year 2 sample. Any new frame hospitals (those in the year 2 frame but not in the year 1 frame) were selected at the rate P_2, using the systematic sampling method described for the 1988 sample selection in Technical Supplement: *Design of the HCUP Nationwide Inpatient Sample, Release 1*.

Stratum-Specific Sampling Rate Decreased ($P_2 < P_1$). Now consider the case where the probability of selection decreased between years 1 and 2. First, hospitals new to the frame were sampled with probability P_2. Second, hospitals previously selected for the year 1 sample (that remained in the year 2 frame) were selected for the year 2 sample with probability $P_2 \div P_1$.

The justification for this second procedure was straightforward. For the year 1 sample hospitals that stayed in the frame, the year 1 sample was viewed as the first stage of a two-stage sampling process. The first stage was carried out at the sampling rate of P_1. The second stage was carried out at the sampling rate of $P_2 \div P_1$. Consequently, the "overall" probability of selection was $P_1 \text{ x } P_2 \div P_1 = P_2$.

Stratum-Specific Sampling Rate Increased ($P_2 > P_1$). The procedures associated with the case in which the probability of selection was increased between year 1 and year 2 were equally straightforward. First, hospitals new to the frame were sampled with probability P_2. Second, hospitals that were selected in year 1 (that remained in the year 2 frame) were selected for the year 2 sample. Third, hospitals that were in the frame for both years 1 and 2, but not selected for the year 1 sample, were selected for the year 2 sample with probability $(P_2\text{-}P_1) \div (1\text{-}P_1)$.

The justification for this sampling rate, $(P_2\text{-}P_1) \div (1\text{-}P_1)$, is somewhat complex. In year 1 certain frame hospitals were included in the sample at the rate P_1. This can also be viewed as having excluded a set of hospitals at the rate $(1\text{-}P_1)$. Likewise, in year 2 it was imperative that each hospital excluded from the year 1 sample be excluded from the year 2 sample at an overall rate of $(1\text{-}P_2)$.

Since $P_2 > P_1$, then $(1\text{-}P_2) < (1\text{-}P_1)$. Therefore, just as was done for the case of $P_2 < P_1$, multistage selection was implemented. However, it was implemented for exclusion rather than inclusion.

Therefore, those hospitals excluded from the year 1 sample were also excluded from the year 2 sample at the rate $S = (1\text{-}P_2) \div (1\text{-}P_1)$. This gave them the desired overall *exclusion* rate of $(1\text{-}P_1) \text{ x } (1\text{-}P_2) \div (1\text{-}P_1) = (1\text{-}P_2)$. Consequently, the *inclusion* rate for these hospitals was set at $1\text{-}S = (P_2\text{-}P_1) \div (1\text{-}P_1)$.

Zero-Weight Hospitals

Beginning in 1993, the NIS samples contain no zero-weight hospitals. For a description of zero-weight hospitals in the 1988-1992 sample, see the Technical Supplement: *Design of the HCUP Nationwide Inpatient Sample, Release 1*.

Ten Percent Subsamples

Two nonoverlapping 10 percent subsamples of discharges were drawn from the NIS file for each year. The subsamples were selected by drawing every tenth discharge starting with two different starting points (randomly selected between 1 and 10). Having a different starting point for each of the two subsamples guaranteed that they would not overlap. Discharges were sampled so that 10 percent of each hospital's discharges in each quarter were selected for each of the subsamples. The two samples can be combined to form a single, generalizable 20 percent subsample of discharges.

FINAL HOSPITAL SAMPLE

The annual numbers of hospitals and discharges in each release of the NIS are shown in Table 50 for both the regular NIS sample and the total sample (which includes zero-weight hospitals for 1988-1992).

Table 50. NIS Hospital Sample

	Regular Sample		Total Sample	
Year	**Number of Hospitals**	**Number of Discharges**	**Number of Hospitals**	**Number of Discharges**
NIS, Release 1				
1988	758	5,242,904	759	5,265,756
1989	875	6,067,667	882	6,110,064
1990	861	6,156,638	871	6,268,515
1991	847	5,984,270	859	6,156,188
1992	838	6,008,001	856	6,195,744
NIS, Release 2				
1993	913	6,538,976	913	6,538,976
NIS, Release 3				
1994	904	6,385,011	904	6,385,011

Table 50. NIS Hospital Sample

Year	Regular Sample		Total Sample	
	Number of Hospitals	Number of Discharges	Number of Hospitals	Number of Discharges
NIS, Release 4				
1995	938	6,714,935	938	6,714,935
NIS, Release 5				
1996	906	6,542,069	906	6,542,069
NIS, Release 6				
1997	1,012	7,148,420	1,012	7,148,420
Total		62,788,891		63,325,678

A more detailed breakdown of the 1997 NIS hospital sample by geographic region is shown in Table 51. For each geographic region, Table 51 shows the number of:

- universe hospitals (Universe),
- frame hospitals (Frame),
- sampled hospitals (Sample),
- target hospitals (Target = 20 percent of the universe), and
- shortfall hospitals (Shortfall = Sample - Target).

Table 51. Number of Hospitals in the Universe, Frame, Regular Sample, Target, and Shortfall by Region, 1997

Region	Universe	Frame	Sample	Target	Shortfall
NE	737	616	154	147	7
MW	1,453	546	302	291	11
S	1,968	553	365	394	-29
W	955	737	191	191	0
Total	5,113	2,452	1,012	1,023	-11

For example, in 1997 the Northeast region contained 737 hospitals in the universe. It also contained 616 hospitals in the frame, of which 154 hospitals were drawn for the sample. This was seven hospitals more than the target sample size of 147.

Table 52 shows the number of hospitals in the universe, frame, and regular sample for each state in the sampling frame for 1997. The difference between the universe and the frame represents the difference in the number of community hospitals in the 1997 AHA Annual Survey of Hospitals and the number of community hospitals for which data were supplied to HCUP in all states except Georgia, Hawaii, Illinois, South Carolina, Tennessee and Missouri.

- The number of hospitals in the Georgia frame is one less than the Georgia universe. One hospital was excluded because of the sampling restrictions stipulated by Georgia.

- The number of hospitals in the Hawaii frame is nine fewer than the Hawaii universe. Six hospitals were excluded because of sampling restrictions stipulated by Hawaii, and three hospitals were not included in the data supplied to HCUP.

- The number of hospitals in the Illinois frame is approximately 55 percent of the hospitals in the Illinois universe in order to comply with the agreement with the data source concerning the restriction on the number of Illinois discharges.

- The number of hospitals in the South Carolina frame is eleven fewer than the South Carolina universe. Six hospitals were excluded because of sampling restrictions stipulated by South Carolina, and five hospitals were not included in the data supplied to HCUP.

- The number of hospitals in the Tennessee frame is 34 fewer than the Tennessee universe. Five hospitals were excluded because of sampling restrictions stipulated by Tennessee, and 29 hospitals were not included in the data supplied to HCUP.

- The number of hospitals in the Missouri frame is 50 fewer than the Missouri universe. Thirty-five hospitals were excluded because they signed release for confidential use only, and 15 hospitals were not included in the data supplied to HCUP.

The number of hospitals in the NIS hospital samples that continue across multiple sample years is shown in Table 53. This table will be of interest to those who may combine Releases 1 through 6 of the NIS. Table 53 shows that longitudinal cohorts that span several years and include 1988 and 1993 are the lowest in number of continuing sample hospitals. For example, if 1988 is taken as a starting year, only 30.7 percent of the 1988 hospital sample continued in the 1997 sample (233 of 758).

Table 52. Number of Hospitals in the Universe, Frame, and Sample for States in the Sampling Frame, 1997

State	Universe	Frame	Sample
AZ	64	62	14
CA	415	411	107
CO	67	66	18
CT	34	32	9
FL	210	198	117
GA	159	158	115
HI	20	11	3

Table 52. Number of Hospitals in the Universe, Frame, and Sample for States in the Sampling Frame, 1997

State	Universe	Frame	Sample
IA	115	115	52
IL	203	112	73
KS	131	120	62
MA	84	73	18
MD	51	51	35
MO	125	75	44
NJ	85	78	19
NY	225	222	56
OR	61	59	16
PA	217	211	52
SC	65	54	34
TN	126	92	64
UT	41	40	13
WA	89	88	20
WI	124	124	71
Total	2,711	2,452	1,012

Table 53. Number of Hospitals and Discharges in Longitudinal Cohorts, 1988-1997

Number of Years	Calendar Years	Longitudinal Regular Sample Hospitals	% of Base Year Sample	Longitudinal Regular Sample Discharges
2	1988-1989	610	80.5	8,492,039
	1989-1990	815	93.1	11,525,749
	1990-1991	802	93.1	11,297,175
	1991-1992	781	92.2	11,272,981
	1992-1993	609	72.7	8,804,638
	1993-1994	693	75.9	10,271,404
	1994-1995	762	84.3	10,747,682
	1995-1996	906	96.6	13,050,676
	1996-1997	741	81.8	10,743,200
3	1988-1990	573	75.6	12,168,677
	1989-1991	763	87.2	16,074,381
	1990-1992	745	86.5	16,085,651
	1991-1993	570	67.3	12,559,421
	1992-1994	540	64.4	11,279,667
	1993-1995	598	65.5	13,241,070
	1994-1996	740	81.9	15,651,230
	1995-1997	741	79.0	16,058,401
4	1988-1991	542	71.5	15,096,807
	1989-1992	709	81.0	20,340,970
	1990-1993	548	63.6	16,023,500
	1991-1994	508	60.0	14,481,319
	1992-1995	464	55.4	12,712,613
	1993-1996	583	63.9	17,203,387
	1994-1997	617	68.3	17,490,946
5	1988-1992	502	66.2	18,106,098
	1989-1993	523	59.8	19,000,777
	1990-1994	490	56.9	17,437,229
	1991-1995	439	51.8	15,405,253
	1992-1996	453	54.1	15,509,564
	1993-1997	485	53.1	17,972,148

Table 53. Number of Hospitals and Discharges in Longitudinal Cohorts, 1988-1997

Number of Years	Calendar Years	Longitudinal Regular Sample Hospitals	% of Base Year Sample	Longitudinal Regular Sample Discharges
6	1988-1993	378	49.9	16,906,818
	1989-1994	471	53.8	19,987,910
	1990-1995	422	49.0	14,817,797
	1991-1996	429	50.6	18,041,571
	1992-1997	379	45.2	15,670,244
7	1988-1994	335	44.2	17,128,064
	1989-1995	408	46.6	19,924,107
	1990-1996	413	48.0	20,293,152
	1991-1997	359	42.4	17,776,471
8	1988-1995	289	38.1	16,658,485
	1989-1996	400	45.7	22,403,308
	1990-1997	344	40.0	19,488,725
9	1988-1996	283	37.3	18,576,353
	1989-1997	334	38.2	21,183,992
10	1988-1997	233	30.7	17,411,298

SAMPLING WEIGHTS

Although the sampling design was simple and straightforward, it is necessary to incorporate sample weights to obtain state and national estimates. Therefore, sample weights were developed separately for hospital- and discharge-level analyses. Three hospital-level weights were developed to weight NIS sample hospitals to the state, frame, and universe. Similarly, three discharge-level weights were developed to weight NIS sample discharges to the state, frame, and universe.

Hospital-Level Sampling Weights

Universe Hospital Weights. Hospital weights to the universe were calculated by post-stratification. For each year, hospitals were stratified on the same variables that were used for sampling: geographic region, urban/rural location, teaching status, bedsize, and control. The strata that were collapsed for sampling were also collapsed for sample weight calculations. Within stratum s, each NIS sample hospital's universe weight was calculated as:

$$W_s(\text{universe}) = N_s(\text{universe}) \div N_s(\text{sample}),$$

where N_s(universe) and N_s(sample) were the number of community hospitals within stratum s in the universe and sample, respectively. Thus, each hospital's universe weight is equal to the number of universe hospitals it represented during that year.

Frame Hospital Weights. Hospital-level sampling weights were also calculated to represent the entire collection of states in the frame using the same post-stratification scheme as described above for the weights to represent the universe. For each year, within stratum s, each NIS sample hospital's frame weight was calculated as:

$$W_s(\text{frame}) = N_s(\text{frame}) \div N_s(\text{sample}).$$

N_s(frame) was the total number of universe community hospitals within stratum s in the states that contributed data to the frame. N_s(sample) was the number of sample hospitals selected for the NIS in stratum s. Thus, each hospital's frame weight is equal to the number of universe hospitals it represented in the frame states during that year.

State Hospital Weights. For each year, a hospital's weight to its state was calculated in a similar fashion. Within each state, strata often had to be collapsed after sample selection for development of weights to ensure a minimum of two sample hospitals within each stratum. For each state and each year, within stratum s, each NIS sample hospital's state weight was calculated as:

$$W_s(\text{state}) = N_s(\text{state}) \div N_s(\text{state sample}).$$

N_s(state) was the number of universe community hospitals in the state within stratum s. N_s(state sample) was the number of hospitals selected for the NIS from that state in stratum s. Thus, each hospital's state weight is equal to the number of hospitals that it represented in its state during that year.

All of these hospital weights can be rescaled if necessary for selected analyses, to sum to the NIS hospital sample size each year.

Discharge-Level Sampling Weights

The calculations for discharge-level sampling weights were very similar to the calculations of hospital-level sampling weights. The discharge weights usually are constant for all discharges within a stratum.

The only exceptions were for strata with sample hospitals that, according to the AHA files, were open for the entire year but contributed less than their full year of data to the NIS. For those hospitals, we *adjusted* the number of observed discharges by a factor 4 ÷ Q, where Q was the number of calendar quarters that the hospital contributed discharges to the NIS. For example, when a sample hospital contributed only two quarters of discharge data to the NIS, the *adjusted* number of discharges was double the observed number.

With that minor adjustment, each discharge weight is essentially equal to the number of reference (universe, frame, or state) discharges that each sampled discharge represented in its stratum. This calculation was possible because the number of total discharges was available for every hospital in the universe from the AHA files. Each universe hospital's AHA discharge total was calculated as the sum of newborns and total facility discharges.

Universe Discharge Weights. Discharge weights to the universe were calculated by post-stratification. Hospitals were stratified just as they were for universe hospital weight calculations. Within stratum s, for hospital i, each NIS sample discharge's universe weight was calculated as:

$$DW_{is}(\text{universe}) = [DN_s(\text{universe}) \div ADN_s(\text{sample})] * (4 \div Q_i),$$

where DN_s(universe) was the number of discharges from community hospitals in the universe within stratum s; ADN_s(sample) was the number of *adjusted* discharges from sample hospitals selected for the NIS; and Q_i was the number of quarters of discharge data contributed by hospital i to the NIS (usually $Q_i = 4$). Thus, each discharge's weight is equal to the number of universe discharges it represented in stratum s during that year.

Frame Discharge Weights. Discharge-level sampling weights were also calculated to represent all discharges from the entire collection of states in the frame using the same post-stratification scheme described above for the discharge weights to represent the universe. For each year, within stratum s, for hospital i, each NIS sample discharge's frame weight was calculated as:

$$W_{is}(\text{frame}) = [DN_s(\text{frame}) \div ADN_s(\text{sample})] * (4 \div Q_i),$$

DN_s(frame) was the number of discharges from all community hospitals in the states that contributed to the frame within stratum s. ADN_s(sample) was the number of *adjusted* discharges from sample hospitals selected for the NIS in stratum s. Q_i was the number of quarters of discharge data contributed by hospital i to the NIS (usually $Q_i = 4$). Thus, each discharge's frame weight is equal to the number of discharges it represented in the frame states during that year.

State Discharge Weights. A discharge's weight to its state was similarly calculated. Strata were collapsed in the same way as they were for the state hospital weights to ensure a minimum of two sample hospitals within each stratum. Within stratum s, for hospital i, each NIS sample discharge's state weight was calculated as:

$$W_{is}(\text{state}) = [DN_s(\text{state}) \div ADN_s(\text{state sample})] * (4 \div Q_i),$$

DN_s(state) was the number of discharges from all community hospitals in the state within stratum s. ADN_s(state sample) was the *adjusted* number of discharges from hospitals selected for the NIS from that state in stratum s. Q_i was the number of quarters of discharge data contributed by hospital i to the NIS (usually $Q_i = 4$). Thus, each discharge's state weight is equal to the number of discharges that it represented in its state during that year.

All of these discharge weights can be rescaled if necessary for selected analyses, to sum to the NIS discharge sample size each year.

Discharge Weights for 10 Percent Subsamples

In the 10 percent subsamples, each discharge had a 10 percent chance of being drawn. Therefore, the discharge weights contained in the Hospital Weights file can be multiplied by 10 for each of the subsamples, or multiplied by 5 for the two subsamples combined.

DATA ANALYSIS

Variance Calculations

It may be important for researchers to calculate a measure of precision for some estimates based on the NIS sample data. Variance estimates must take into account both the sampling design and the form of the statistic. The sampling design was a stratified, single-stage cluster sample. A stratified random sample of hospitals (clusters) were drawn and then *all* discharges were included from each selected hospital.

If hospitals inside the frame were similar to hospitals outside the frame, the sample hospitals can be treated as if they were randomly selected from the entire universe of hospitals within each stratum. Standard formulas for a stratified, single-stage cluster sampling without replacement could be used to calculate statistics and their variances in most applications.

A multitude of statistics can be estimated from the NIS data. Several computer programs are listed below that calculate statistics and their variances from sample survey data. Some of these programs use general methods of variance calculations (e.g., the jackknife and balanced half-sample replications) that take into account the sampling design. However, it may be desirable to calculate variances using formulas specifically developed for some statistics.

In most cases, computer programs are readily available to perform these calculations. For instance, OSIRIS IV, developed at the University of Michigan, and SUDAAN, developed at the Research Triangle Institute, do calculations for numerous statistics arising from the stratified, single-stage cluster sampling design. An example of using SUDAAN to calculate variances in the NIS is presented in Technical Supplement: *Calculating Variances Using Data from the HCUP Nationwide Inpatient Sample*.[3]

These variance calculations are based on finite-sample theory, which is an appropriate method for obtaining cross-sectional, nationwide estimates of outcomes. According to finite-sample theory, the intent of the estimation process is to obtain estimates that are precise representations of the nationwide population at a specific point in time. In the context of the NIS, any estimates that attempt to accurately describe characteristics (such as expenditure and utilization patterns or hospital market factors) and interrelationships among characteristics of hospitals and discharges during a specific year from 1988 to 1997 should be governed by finite-sample theory.

Alternatively, in the study of hypothetical population outcomes not limited to a specific point in time, analysts may be less interested in specific characteristics from the finite population (and time period) from which the *sample* was drawn, than they are in hypothetical characteristics of a conceptual "superpopulation" from which any particular finite *population* in a given year might have been drawn. According to this superpopulation model, the nationwide population in a given year is only a snapshot in time of the possible interrelationships among hospital, market, and discharge characteristics. In a given year, all possible interactions between such characteristics may not have been observed, but analysts may wish to predict or simulate interrelationships that may occur in the future.

Under the finite-population model, the variances of estimates approach zero as the sampling fraction approaches one, since the population is defined at that point in time, and because the estimate is for a characteristic as it existed at the time of sampling. This is in contrast to the superpopulation model, which adopts a stochastic viewpoint rather than a deterministic viewpoint. That is, the nationwide population in a particular year is viewed as a random sample of some underlying superpopulation over time.

Different methods are used for calculating variances under the two sample theories. Under the superpopulation (stochastic) model, procedures (such as those described by Potthoff, Woodbury, and Manton[4]) have been developed to draw inferences using weights from complex samples. In this context, the survey weights are not used to weight the sampled cases to the universe, because the universe is conceptually infinite in size. Instead, these weights are used to produce unbiased estimates of parameters that govern the superpopulation.

In summary, the choice of an appropriate method for calculating variances for nationwide estimates depends on the type of measure and the intent of the estimation process.

Computer Software for Variance Calculations

The hospital weights will be useful for producing hospital-level statistics for analyses that use the *hospital* as the unit of analysis, and the discharge weights will be useful for producing discharge-level statistics for analyses that use the *discharge* as the unit of analysis. These would be used to weight the sample data in estimating population statistics.

Several statistical programming packages allow weighted analyses.[5] For example, nearly all SAS (Statistical Analysis System) procedures incorporate weights.

In addition, several statistical analysis programs have been developed that specifically calculate statistics and their standard errors from survey data. For an excellent review of such programs, visit the following web site: http://www.fas.harvard.edu/~stats/survey-soft/.

The NIS database includes a Hospital Weights file with variables required by these programs to calculate finite population statistics. In addition to the sample weights described earlier, hospital identifiers (PSUs), stratification variables, and stratum-specific totals for the numbers of discharges and hospitals are included so that finite-population corrections (FPCs) can be applied to variance estimates.

In addition to these subroutines, standard errors can be estimated by validation and cross-validation techniques. Given that a very large number of observations will be available for most analyses, it may be feasible to set aside a part of the data for validation purposes. Standard errors and confidence intervals can then be calculated from the validation data. If the analytical file is too small to set aside a large validation sample, cross-validation techniques may be used.

For example, tenfold cross-validation would split the data into ten equal-sized subsets. The estimation would take place in ten iterations. At each iteration, the outcome of interest is predicted for one-tenth of the observations by an estimate based on a model fit to the other nine-tenths of the observations. Unbiased estimates of error variance are then obtained by comparing the actual values to the predicted values obtained in this manner.

Finally, it should be noted that a large array of hospital-level variables are available for the entire universe of hospitals, including those outside the sampling frame. For instance, the variables from the AHA surveys and from the Medicare Cost Reports are available for nearly all hospitals. To the extent that hospital-level outcomes correlate with these variables, they may be used to sharpen regional and nationwide estimates.

As a simple example, each hospital's number of C-sections would be correlated with their total number of deliveries. The number of C-sections must be obtained from discharge data, but the number of deliveries is available from AHA data. Thus, if a regression can be fit predicting

C-sections from deliveries based on the NIS data, that regression can then be used to obtain hospital-specific estimates of the number of C-sections for all hospitals in the universe.

Longitudinal Analyses

As previously shown in Table 53, hospitals that continue in the NIS for multiple consecutive years are a subset of the hospitals in the NIS for any one of those years. Consequently, longitudinal analyses of hospital-level outcomes may be biased if they are based on any subset of NIS hospitals limited to continuous NIS membership. In particular, such subsets would tend to contain fewer hospitals that opened, closed, split, merged, or changed strata. Further, the sample weights were developed as annual, cross-sectional weights rather than longitudinal weights. Therefore, different weights might be required, depending on the statistical methods employed by the analyst.

One approach to consider in hospital-level longitudinal analyses is to use repeated-measure models that allow hospitals to have missing values for some years. However, the data are not actually missing for some hospitals, such as those that closed during the study period. In any case, the analyses may be more efficient (e.g., produce more precise estimates) if they account for the potential correlation between repeated measures on the same hospital over time, yet incorporate data from all hospitals in the sample during the study period.

Discharge Subsamples

The two nonoverlapping 10 percent subsamples of discharges were drawn from the NIS file for each year for several reasons pertaining to data analysis. One reason for creating the subsamples was to reduce processing costs for selected studies that will not require the entire NIS. Another reason is that the two subsamples may be used to validate models and obtain unbiased estimates of standard errors. That is, one subsample may be used to estimate statistical models, and the other subsample may be used to test the fit of those models on new data. This is a very important analytical step, particularly in exploratory studies, where one runs the risk of fitting noise.

For example, it is well known that the percentage of variance explained by a regression, R^2, is generally overestimated by the data used to fit a model. The regression model could be estimated from the first subsample and then applied to the second subsample. The squared correlation between the actual and predicted value in the second subsample is an unbiased estimate of the model's true explanatory power when applied to new data.

ENDNOTES

1. Most AHA surveys do not cover a January-to-December calendar year. The number of hospitals for 1988-1991 are based on the HCUP calendar-year version of the AHA Annual Survey files. To create a calendar-year reporting period, data from the AHA surveys must be apportioned in some manner across calendar years. Survey responses were converted to calendar-year periods for 1988-1991 by merging data from adjacent survey years. The number of hospitals for 1992-1997 are based on the AHA Annual Survey files.

2. Coffey, R. and D. Farley (1988, July). *HCUP-2 Project Overview,* (DHHS Publication No. (PHS) 88-3428. Hospital Studies Program Research Note 10, National Center for Health Services Research and Health Care Technology Assessment, Rockville, MD: Public Health Service.

3. Duffy, S.Q. and J.P. Sommers (1996, March). *Calculating Variances Using Data from the HCUP Nationwide Inpatient Sample*. Rockville, MD: Agency for Health Care Policy and Research.

4. Potthoff, R.F., M.A. Woodbury, and K.G. Manton (1992). "Equivalent Sample Size" and "Equivalent Degrees of Freedom" Refinements for Inference Using Survey Weights Under Superpopulation Models. *Journal of the American Statistical Association*, Vol. 87, 383-396.

5. Carlson, B.L., A.E. Johnson, and S.B. Cohen (1993). An Evaluation of the Use of Personal Computers for Variance Estimation with Complex Survey Data. *Journal of Official Statistics*, Vol. 9, No. 4, 795-814.

TECHNICAL SUPPLEMENT 11: CALCULATING VARIANCES USING DATA FROM THE HCUP NATIONWIDE INPATIENT SAMPLE

Sarah Q. Duffy, Ph.D. and John P. Sommers, Ph.D.
Agency for Health Care Policy and Research

INTRODUCTION

The Nationwide Inpatient Sample (NIS), Release 1 database contains all discharges from hospitals that were selected without replacement according to a stratified probability sample design from a frame that includes hospitals from 8 states for 1988 and 11 states for 1989-1992. Release 2 and Release 3 include data from 17 states for 1993 and 1994, respectively, and Release 4 and Release 5 include data from 19 states for 1995 and 1996, and Release 6 includes data from 22 states in 1997. Failure to account for this sample design when computing statistics will cause variances to be estimated incorrectly. This document states the problem and gives an example of how one readily available complex survey design package, the Survey Data Analysis Software System, or SUDAAN, can be used to estimate variances while accounting for the sample design of the NIS. The reader should be prepared to consult the SUDAAN documentation as necessary.

Due to the correlation between observations caused by the same hospitals appearing in the NIS data across years, it is difficult to calculate standard errors when multiple years of data are pooled in one analytic dataset. The methods described in this paper are appropriate for calculating variances using one year of NIS data at a time.

BACKGROUND

Variances Based on Simple Random Sampling

Many popular statistical packages, such as SAS and SPSS, use the following formula based on simple random sampling to calculate an estimate for the sample variance:

$$\hat{o}^2 = \frac{\sum (y_{hij} - \bar{y})^2}{n - 1}$$

where:

$\hat{o}^2$ = variance estimate

ó = the standard deviation

y_{hij} = the value of variable *y* for the *jth* sample discharge in the *ith* sample hospital in the *hth* stratum

$\bar{y}$ = the grand mean of the variable y, and

n = the number of observations in the sample.

Variances Based on the NIS

Since the NIS is not a simple random sample, it requires a different variance formula. The NIS sample design has several characteristics that require modification of the variance formula: sample weights, two-stage sampling from a finite population, and stratification. Complex survey design packages such as SUDAAN (descriptive statistics), SURREGER (ordinary least squares regression), and RTILOGIT (logistic regression) allow these characteristics to be incorporated into variance estimation.[1]

Variance formulas appropriate for the NIS data contain weights, components for the two stages of sampling, and factors to correct for the proportion of the frame included in the sample at each level (finite population correction factors). For example, define the weighted sum, Y,

$$Y = \sum_{h}^{H} \sum_{i}^{n_h} \sum_{j}^{n_{hi}} w_{hij} y_{hij}$$

where:

y_{hij} = the value of a variable *y* for the *jth* sample discharge in the *ith* sample hospital in the *hth* stratum, as above

w_{hij} = a set of weights or any other constants over the set of sample discharges, hospitals, and strata

n_{hi} = the number of discharges in the *ith* sample hospital in the *hth* stratum

n_h = the number of sample hospitals in the *hth* stratum and

H = the number of strata.

Then the estimate of the variance of Y from the sample, $\hat{o}_Y^2$, is

$$\hat{o}_Y^2 = \sum_{h}^{H} (1-f_h) n_h S_h^2 + \sum_{h}^{H} f_h \sum_{i}^{n_{hi}} (1-f_{hi}) n_{hi} S_{hi}^2 \qquad (1)$$

where:

$\hat{o}$ = the standard deviation of Y

f_h = the proportion of the total number of hospitals in the *hth* stratum selected into the sample, i.e., the first stage sampling rate in the *hth* stratum. (This is simply the number of hospitals from stratum h in the sample divided by the total number of hospitals in stratum h on the frame.)

f_{hi} = the proportion of the discharges in the sample from the *ith* sample hospital in the *hth* stratum, i.e., the second stage sampling rate in the *hith* hospital.[2] (This is simply the number of discharges from hospital *i* in stratum *h* in the sample divided by the total number of discharges in hospital *i* and stratum *h.*)

S_h^2 = the component for the first stage of sampling, the overall variation due to variation between hospitals within strata

$$= \frac{\sum_{i}^{n_h} \left(\sum_{j}^{n_{hi}} w_{hij} y_{hij} - \frac{\sum_{i}^{n_h} \sum_{j}^{n_{hi}} w_{hij} y_{hij}}{n_h} \right)^2}{(n_h - 1)} \text{ and}$$

S_{hi}^2 = a portion of the component for the second stage of sampling, the overall variation of discharges within hospital for the *hith* hospital.

$$= \frac{\sum_{j}^{n_{hi}} \left(w_{hij} y_{hij} - \frac{\sum_{j}^{n_{hi}} w_{hij} y_{hij}}{n_{hi}} \right)^2}{n_{hi} - 1} \; .$$

SUDAAN uses variance formulas similar to (1) in its DESCRIPT procedure to calculate a large number of simple statistics, such as estimates of means and totals. Variances of more complex estimates, such as ratios of two different random variables, require formulas that include contributions from the variances of each of the random variables. These can be calculated using the SUDAAN RATIO procedure.

Because the states included in the NIS frame were not selected randomly, the error associated with statistical estimates derived from NIS data actually contains two parts:

a. variance formula (1), the error due to sampling from the selected states, and
b. the bias from not using the entire U.S. as the sampling frame.

Part b, the bias, cannot be calculated directly. When possible, estimates based on the NIS should be checked against national benchmarks, such as Medicare data or data from the National Hospital Discharge Survey, to determine the appropriateness of the NIS for specific analyses.

AN EXAMPLE: CALCULATING MEAN LENGTH OF STAY FOR AMI PATIENTS

SUDAAN, available for both the PC and the mainframe, can be used to calculate variance estimates for simple descriptive statistics (means, percentiles, ratios) and cross tabulations.[3] For example, suppose an analyst wants to calculate a nationally weighted mean length of stay (LOS) and its variance for all patients in the NIS with acute myocardial infarction (AMI) as a principal diagnosis, using the SUDAAN DESCRIPT procedure.

Description of SUDAAN Code

The SUDAAN code would be as follows:

```
PROC DESCRIPT  DATA=analysis file  FILETYPE=analysis file type  DESIGN=WOR;

WEIGHT DISCWT_U;

NEST STRATUM HOSPID/MISSUNIT;

TOTCNT N_HOSP_F _ZERO_;

SAMCNT S_HOSP_U TOTDSCHG;

VAR LOS;

PRINT MEAN SEMEAN;

OUTPUT MEAN SEMEAN/ FILETYPE = ASCII FILENAME = "output file";
```

Details:

PROC DESCRIPT

DATA = the name of the analysis file. In this example, the analysis file consists of all discharges in the NIS with an AMI coded as a principal diagnosis. The input analysis file must always be sorted by the variables specified in the NEST statement (see below), in this case STRATUM and HOSPID.

FILETYPE = the format of the analysis file. SUDAAN will read SAS and ASCII files.

DESIGN = the sample design. For the NIS, WOR (without replacement) is the appropriate choice. See the SUDAAN documentation for details.

Requests for statistics other than means, such as quantiles, would be included in the PROC DESCRIPT statement as well. Since mean is the default statistic, it is not necessary to specify it.

WEIGHT

Specifies the weighting variable. It is a required statement. The NIS variable DISCWT_U is the weight for discharges in the NIS, and must be merged onto the analysis file from the NIS, Release 1 Hospital Weights file. (DISCWT_U has been merged to Inpatient Stay Core File A for NIS, Release 2, Release 3, Release 4 and Release 5.) To get unweighted statistics, simply create a variable that is equal to 1 for all observations and specify that variable in the WEIGHT statement.

NEST

Specifies the variables corresponding to the levels in the sampling design. It is a required statement. In the NIS, hospitals were selected from strata, which are identified by the NIS variable STRATUM, and discharges were selected from hospitals, which are identified by the NIS variable HOSPID. The analysis file must be sorted by the variables in this statement.

TOTCNT

Specifies the population counts at each stage of the sampling design for which without replacement sampling is assumed. It is a required statement. The NIS variable N_HOSP_F contains the number of hospitals on the frame in the stratum, and must be merged onto the analysis file from the NIS hospital-level file.[4]

ZERO is a variable available to all SUDAAN procedures that is used to identify stages for which no sample selection took place. It prompts SUDAAN not to calculate the corresponding variance component. It is used in this example for the hospital-level counts because there was no sampling at this level - all AMI discharges are included in the analysis file.[5] This makes each f_{hi} equal to one in formula (1) effectively zeroing the second term. Hence, the name _ZERO_.

SAMCNT

Specifies the sample counts at each stage of the sample design for which without replacement sampling is assumed. It is an optional statement. The NIS variable S_HOSP_U, again from the hospital file, contains the number of hospitals sampled in each stratum. The NIS variable TOTDSCHG contains the number of discharges in the hospital, which must be specified even though _ZERO_ is specified.

SUDAAN double checks the accuracy of the variables specified in the SAMCNT statement by counting the number of observations in the file at each level. It will do this whether or not the sample count variables are specified. In this case, since SUDAAN's count will be the correct value, it may make most sense to omit the SAMCNT statement from program. However, there are applications for which SUDAAN's count will be incorrect. See the section "Using the NIS Subsamples", below, for an example of such an application.

VAR

Specifies the variables for which statistics are to be calculated, in this case LOS. Multiple variables may be included on the VAR statement, but they must all be of one type, continuous or discrete.

PRINT

Specifies that the mean (MEAN) and its standard error (SEMEAN) be printed. SEMEAN, the standard error of the mean, is equivalent to the $\hat{\sigma}$ referred to earlier and should be used when calculating Z scores and other tests of significance.

OUTPUT

Specifies the ASCII output files to which the results are to be read and requests that MEAN and SEMEAN be included on the files. SUDAAN creates 5 files, each of which has as its root name the name specified on the statement. Details may be found in the SUDAAN documentation.

Steps in Computation

To demonstrate the effect of using SUDAAN, standard errors were calculated using both the above program and SAS PROC MEANS. This involved:

1. Pulling all discharges from the 1992 NIS with DCCHPR1 = 100 (acute myocardial infarction).

2. Merging the resulting file to the NIS, Release 1 Hospital Weights File.

3. Downloading the data to a secure PC environment.

4. Running PC SUDAAN PROC DESCRIPT to get weighted national estimates of

    a. the number of AMI patients, and
    b. mean length of stay of AMI patients and its standard error.

5. Running PC SAS to get both weighted and unweighted estimates of the variables mentioned in step 4. The SAS weighted estimates were computed two ways:

    a. using the WEIGHT Statement with the VARDEF=WEIGHT option. When using the WEIGHT statement to get counts, the analyst must specify a variable equal to 1 for each discharge and request a SUM for that variable.
    b. using the FREQ statement. Since the weight variables in the NIS are not integers, using the FREQ statement will underestimate counts, as the results below reveal. This is because SAS uses only the integer portion of variables specified in the FREQ statement.

Results of Computation

The results, displayed in Table 54, reveal that accounting for the sample design affects the estimated variances. In this example, accounting for the sample design resulted in lower variances. The standard error calculated by SUDAAN is only 50% as large as the one calculated using SAS PROC MEANS with the FREQ statement, which as noted above gives incorrect count estimates as well, and less than one quarter the size of that estimated when using SAS PROC MEANS either unweighted or with the WEIGHT statement. Calculating the variances using SUDAAN or other complex survey design package will often result in lower variances from NIS data because of the finite correction factor, but this is by no means guaranteed.

Table 54. Weighted and Unweighted Estimates of Counts, and Average Length of Stay and its Standard Error for 1992 AMI Discharges, HCUP National Inpatient Sample, Release 1

Variable	Weighted PC SUDAAN	Weighted PC SAS Weight Statement[2]	Weighted PC SAS, FREQ Statement[3]	Unweighted PC SAS
Count of AMIs[1]	688,054	688,054	625,605	119,121
Average Length of Stay (Standard Error)	7.88 (0.05)	7.88 (0.21)	7.87 (0.10)	8.01 (0.23)

Notes:

1. AMI discharges are those with DCCHPR1 = 100.
2. Run with the VARDEF = WEIGHT statement. When a WEIGHT statement is specified, the reported N is equal to the unweighted sum. To find the weighted sum the analyst must create a variable that equals 1 for each observation and request SUM on that variable.
3. PROC FREQ results in a lower weighted count because it only uses the integer portion of the weights. As the results reveal, it should not be used to compute weighted estimates from the NIS data.

USING THE NIS SUBSAMPLES

SUDAAN can estimate variances using data from a NIS 10% subsample, but the program must be modified in two places: the WEIGHT statement and the TOTCNT statement. The 10% subsample contains the same hospitals as the full NIS sample, but has only 10% of their discharges rather than the 100% contained in the NIS.

WEIGHT

When using the 10% sample to get weighted estimates, multiply the variable DISCWT_U by 10, and specify the resulting variable on the WEIGHT statement.

TOTCNT

The TOTCNT statement must be modified because SUDAAN counts the observations in the sample at each level and uses the result instead of whatever is specified in the SAMCNT statement. For example, suppose an analyst wants to calculate mean LOS for AMI patients as above, but wants to use a 10% sample. Suppose the analyst specifies N_HOSP_F in the TOTCNT statement, as above, for the number of hospitals on the frame, but specifies TOTDSCHG for the total number of discharges for each hospital. The analyst would then specify S_HOSP_U for the number of sample hospitals in the stratum, along with a newly created variable that contained the number of discharges in the 10% sample from each hospital, .10*TOTDSCHG.

However, as mentioned above, SUDAAN will double check the counts of the variables specified in the SAMCNT statement. When it counts the hospitals, it will determine the correct number. But when it counts the discharges on the analysis file, it will find far fewer than 10% of each hospital's actual discharges because the analysis file contains only AMIs. SUDAAN will use the count of AMI discharges instead of all discharges in the 10% sample, which will cause it to underestimate the finite correction f_{hi}, thus overstating the variance.

There are at least two ways to modify the TOTCNT statement to avoid this problem when using the 10% sample:

1. The first is to specify _MINUS1_ instead of TOTDSCHG in the TOTCNT statement. This essentially tricks SUDAAN into calculating the variance without implementing the finite correction factor for this component, since specifying _MINUS1_ signals to SUDAAN that sampling at this level was with replacement. Since the finite correction factor for a 10% sample would be .9, calculating the variance this way will lead to a slight overestimate of the variance.

2. A more precise way would be to create a variable equal to the total number of AMIs in the hospital and use that in the TOTCNT statement instead of the total number of discharges in the hospital. This could be estimated as ten times the number of AMI discharges from the 10% sample or the actual number of AMI discharges from the full NIS sample file.

When SUDAAN counts the discharges in the analysis file under either of these approaches and takes the ratio of that number to the variable specified in the TOTCNT statement, it will get about .1, the correct value for f_{hi}.

CALCULATING VALUES FOR SUBSETS OF THE NIS

The SUBGROUP and LEVELS statements from SUDAAN can be used to calculate values for subsets of the entire population using either the full NIS or any NIS subsample. In either case, the entire data set is used. As an example, suppose an analyst desires values for male and female discharges. The analyst can use a variable SEX, where SEX = 1, if the discharge is a male, and SEX = 2, if female. To produce the proper code, the analyst modifies the code given in Section III, if using the full NIS, or Section IV, if using a NIS subsample. The modifications are:

1. Remove the SAMCNT statement (SUDAAN counts the cases) and

2. Add the statements:
   ```
   SUBGROUP SEX;
   LEVELS 2;
   ```
   to designate the partitioning variable and the numbers of partitions.

One can calculate values for more than two subsets in a single step. For example, if an analyst desires m partitions for a variable, the analyst creates variable P with values 1, 2 ...m, which partition the data set into the desired subsets (SUDAAN specifies that only the integers 1, 2, ...m can be used to create the m subsets). The statements:

```
SUBGROUP P;
LEVELS m;
```

are used to denote the partitioning variable and the number of partitions.

More complex partitioning using crosses of multiple variables is also allowed. Details may be found in SUDAAN documentation.

COMPUTER RESOURCE REQUIREMENTS

SUDAAN is an expensive package to run on a mainframe, especially on samples as large as those generated by the NIS, so it may be wise to consider using the PC version of SUDAAN whenever possible. Although resource requirements will vary across applications, a program using SUDAAN to calculate 500 means and their variances for two variables from an analysis file that contained 625,000 observations took 10 minutes on a 100 Mhz Pentium. The program from the example above that calculated the mean length of stay and its standard error for AMI patients took about 5 minutes on the same machine. Files can be manipulated on the mainframe and then downloaded in either ASCII or SAS format to the PC, where they can be read by SUDAAN.

ENDNOTES

1. See Carlson, B. L., A. E. Johnson, and S. B. Cohen, 1993, "An Evaluation of the Use of Personal Computers for Variance Estimation with Complex Survey Data", *Journal of Official Statistics*, 9(4): 795-814 for additional information on these and other complex survey data packages.

2. These finite correction factors $(1\text{-}f_h)$ and $(1\text{-}f_{hi})$ reduce the variance as the sample becomes a larger portion of the frame. This is intuitively plausible since the variance would be zero if the entire frame were included in the sample.

3. SUDAAN may be purchased from the Research Triangle Institute, Research Triangle Park, NC. See Carlson *et al., op. cit.* for sources for other complex survey design packages.

4. The SUDAAN documentation states that the universe count should appear in the TOTCNT statement. That is because in many surveys, the frame *is* the universe. In the NIS the frame is not the universe of all hospitals U.S., as mentioned above, so the appropriate variable for the TOTCNT statement is the count of hospitals in the frame in each stratum.

5. Note that there are approximately 10 to 15 hospitals in each year of the NIS that provide less than a full year's worth of data. For those hospitals the analyst would not have all AMI discharges. However, the effect on the estimated variance is small enough to ignore.

TECHNICAL SUPPLEMENT 12:
FILE COMPOSITION FOR THE HCUP NATIONWIDE INPATIENT SAMPLE

OVERVIEW

The Healthcare Cost and Utilization Project (HCUP) Nationwide Inpatient Sample (NIS) is designed to be a 20 percent sample of U.S. community hospitals, as defined by the American Hospital Association (AHA). The AHA defines community hospitals as "all nonfederal, short-term, general and other specialty hospitals, excluding hospital units of institutions." The HCUP sample is a stratified probability sample of hospitals in the frame, with sampling probabilities proportional to the number of U.S. community hospitals in each stratum. The frame is limited by the availability of data.

The hospital universe is defined using the AHA Annual Survey of Hospitals. This universe of hospitals is divided into strata using five hospital characteristics: ownership/control, bedsize, teaching status, rural/urban location, and U.S. region. Hospitals from HCUP participating states (the sampling frame) are selected to represent these strata, and all discharges from sampled hospitals are included in the database. To allow for the production of national estimates, both hospital and discharge weights are provided along with information necessary to calculate the variance of estimates. The weights were developed from the same AHA-defined characteristics that define HCUP sampling strata.

States in the NIS

The NIS is comprised of selected states that have agreed to provide the project with all-payer data on hospital inpatient stays. Different releases of the NIS span different years and include different numbers of states:

- NIS, Release 6 contains 1997 data from 22 states;
- NIS, Release 5 contains 1996 data from 19 states;
- NIS, Release 4 contains 1995 data from 19 states;
- NIS, Release 3 contains 1994 data from 17 states;
- NIS, Release 2 contains 1993 data from 17 states; and
- NIS, Release 1 covers the years 1988 through 1992 and is drawn from 11 states (only 8 states are included 1988).

The NIS contains all discharges from hospitals sampled from these states.

NIS Data Files

There are two different types of NIS data:

- Data on inpatient stays; and
- Data on hospitals, in the Hospital Weights file.

There are three main collections of NIS inpatient data:

- 100% of inpatient records for each sampled hospital; and
- Two non-overlapping 10% subsamples of inpatient records from all NIS hospitals.

Inpatient data elements include linkage elements, patient demographics, clinical information, and payment information. For more information on the structure of the NIS Inpatient Stay files, refer to the release-specific NIS Documentation.

The NIS Hospital Weights file contains one observation per year for each hospital included in the NIS. This file contains data elements for linkage, strata definitions, and sample weights. Strata variables are based on information from the AHA Annual Survey of Hospitals. Sample weights were developed separately for hospital- and discharge-level analyses for each year. Three hospital-level weights were developed to weight NIS hospitals to the state, frame, and universe. Likewise, three discharge-level weights were developed to weight NIS discharges to the state, frame, and universe. When linked with the NIS Inpatient Stay file by the HCUP hospital identifier (HOSPID), the Hospital Weights file provides all the data elements required to produce national estimates, including the variance of estimates.

For detailed information about the development and use of discharge and hospital weights, see the release-specific Technical Supplements on *Design of the HCUP Nationwide Inpatient Sample*.

HCUP CRITERIA

Criteria for Including Hospitals in HCUP

The American Hospital Association (AHA) Annual Survey definition of a community hospital is used to determine which facilities are eligible for inclusion in the HCUP database. If the AHA Annual Survey considers a hospital to be a community hospital, then all of its discharges are eligible for inclusion in the HCUP sample.

The AHA Annual Survey definition of a hospital may not always coincide with the definition of a hospital used by data sources. Specific examples of discrepancies include:

- If a data source reports inpatient data for two or more separate facilities which are considered by the AHA to be a single hospital, HCUP treats them as a single hospital.

- If a data source reports inpatient data from a hospital that cannot be identified in the AHA Annual Survey, that hospital is excluded from the HCUP database.

Federal and Veterans hospitals are excluded from the HCUP database because HCUP data sources do not consistently collect information on these hospitals.

Definition of a Community Hospital

The AHA Annual Survey definition of a community hospital includes nonfederal short-term hospitals whose facilities are available to the public. Short-term is defined as hospitals with an average length of stay less than 30 days. Both general and specialty hospitals (e.g., obstetrics and gynecology, rehabilitation, orthopedics, and eye, ear, nose and throat) are included. There are some hospitals for which the average length of stay for records in the HCUP database is greater than 30 days.

Opened and Closed Hospitals

Hospital openings and closures may be reflected at different times in the supplied inpatient data and the AHA Annual Survey.

Openings. A hospital is included in the HCUP database only when the hospital is recognized by the AHA Annual Survey for that year. This means that inpatient data received from a data source for an opening hospital are excluded from the HCUP database until the hospital is recognized by the AHA Annual Survey. The lag between hospital openings and recognition in the AHA Annual Survey may be more than one year.

Closures. A hospital included in the HCUP database continues to be included if there are inpatient data supplied by the source, even if the AHA Annual Survey considers the hospital to have closed. This means that inpatient data will continue to be included in the HCUP database after the AHA Annual Survey ceases to recognize the hospital, unless there is strong evidence that the hospital has ceased to be a community hospital.

Inclusion of Stays in Special Units

Hospitals may vary in their reporting of discharges from special units (e.g., psychiatric, rehabilitation, long-term care). If information about such reporting is available, it is documented under File Composition by State. No attempt has been made to delete records from special units within hospitals.

"SPECIAL" HOSPITALS WITH ZERO WEIGHTS

To allow for longitudinal analysis of special events such as hospital closures, mergers, and splits, the sample is adjusted to keep these "special" hospitals in the database over time. When hospitals are kept in the database solely because they are "special," zero weights are associated with them (i.e., these hospitals will not be counted in nationally weighted estimates).

Zero-weight hospitals are included in the 1988-1992 data for NIS, Release 1.

Because relatively few hospitals were affected and the complexity of including these hospitals entailed considerable processing burden and costs, no zero-weight hospitals are included after 1992 in NIS Release 2, 3, 4, 5 and 6.

STATES IN THE NIS

The following section lists all states participating in the NIS and provides details about the sources of the data, inclusion of hospital stays in special units, exclusion of ambulatory surgery records, and special precautions required by some states for maintaining confidentiality of hospitals.

Arizona

The HCUP Arizona files were constructed from the Arizona Hospital Inpatient Database from the Cost Reporting and Review Section of the Arizona Department of Health Services. Arizona supplied discharge abstract data for inpatient stays in acute care and rehabilitation hospitals with more than 50 beds. Some community hospitals, as defined by the AHA Annual Survey of Hospitals, may not be included in the HCUP Inpatient Databases because their data were not provided by the data source. In 1997, data from 3% of the community hospitals in Arizona were not received.

Arizona data are included in HCUP beginning in 1989.

Inclusion of Stays in Special Units. The source documentation supplied by Arizona does not indicate whether stays in special units within the hospital (e.g., psychiatric, rehabilitation, long-term care) are included.

California

The HCUP California files were constructed from the confidential files received from the Office of Statewide Health Planning and Development (OSHPD). California supplied discharge abstract data for inpatient stays in general acute care hospitals, acute psychiatric hospitals, chemical dependency recovery hospitals, psychiatric health facilities, and state-operated hospitals. Some community hospitals, as defined by the AHA Annual Survey of Hospitals, may not be included in the HCUP Inpatient Databases because their data were not provided by the data source. In 1997, data from 1% of the community hospitals in California were not received. California excluded inpatient stays that, after processing by OSHPD, did not contain a complete and "in-range" admission date or discharge date. California also excluded inpatient stays that had an unknown or missing date of birth.

California data are included in HCUP beginning in 1988.

Inclusion of Stays in Special Units. Included with the general acute care stays in community hospitals are stays in skilled nursing, intermediate care, rehabilitation, alcohol/chemical dependency treatment, and psychiatric units. Stays in these different types of units can be identified by the first digit of the source hospital identifier (DSHOSPID):

0	=	Type of unit unknown (beginning in 1996)
1	=	General acute care
2	=	Not a valid code
3	=	Skilled nursing and intermediate care (long term care)
4	=	Psychiatric care
5	=	Alcohol/chemical dependency recovery treatment
6	=	Acute physical medicine rehabilitation care.

The reliability of this indicator for the type of care depends on how it was assigned.

Prior to 1995. The type of care was assigned by California based on the hospital's licensed units and the proportion of records in a batch of submitted records that fall into each Major Diagnostic Category (MDC). Hospitals were permitted to submit discharge records in one of two ways: submit separate batches of records for each type of care OR bundle records for all types of care into a single submission. How a hospital submitted its records to California determined the accuracy of the type of care indicated in the first digit of DSHOSPID. Consider a hospital which is licensed for more than one type of care:

- If the hospital submitted one batch of records per type of care, then the distribution of each batch of discharges into MDCs would clearly indicate the type of care (acute, psychiatric, etc.). The data source could then accurately assign the first digit of DSHOSPID.

- If the same hospital submitted all of its records in one batch, then the distribution of discharges into MDCs would be a mixture of acute and other types of care. The first digit of DSHOSPID would be set to "general acute care" (value = 1) on all records and would not distinguish the types of care.

Prior to 1995, most hospitals submitted only one batch of records to California which meant that the type of care indicated in the first digit of DSHOSPID did not distinguish among types of care.

Beginning in 1995. Hospitals were required to assign type of care codes to individual records for certain discharges. These discharges included:

- general acute care (value = 1),
- skilled nursing and intermediate care (value = 3), and
- rehabilitation care (value = 6).

For discharges from facilities licensed as psychiatric care (value = 4) or alcohol/chemical dependency recovery treatment (value = 5), California continued to assign the type of care code to all discharges from the facility.

Colorado

The HCUP Colorado files were constructed from the Discharge Data Program (DDP) files. The Colorado Health and Hospital Association supplied discharge abstract data from Colorado acute care hospitals, including swing beds and distinct part units. Some community hospitals, as defined by the AHA Annual Survey of Hospitals, may not be included in the HCUP Inpatient Databases because their data were not provided by the data source. In 1997, data from 1.5% of the community hospitals in Colorado were not received.

Colorado data are included in HCUP beginning in 1988.

From 1988 to 1990, abstracts for all ambulatory surgeries were also supplied in the source files, but these were excluded from the HCUP inpatient database, as described below.

Starting in 1991, Colorado supplied inpatient and ambulatory surgery records in separate files.

Inclusion of Stays in Special Units. The Colorado Health and Hospital Association does not require hospitals to submit information from their SNFs and ICFs, but no attempt has been made to verify their exclusion.

Exclusion of Ambulatory Surgery Records. For 1988 through 1990, the data source supplied a mixture of inpatient and ambulatory surgery records distinguished by a record type indicator. Only the inpatient discharges were retained in the HCUP files. The table below explains how the inpatient discharges were identified.

Table 55. How Inpatient Records Were Identified in Colorado Data

Record Type	Value of Record Type Indicator on Discharge Abstract	Inclusion in HCUP Data
Inpatient	1	Include
Ambulatory surgery	2	Exclude
Unknown	0	Exclude if all of the following conditions are true (i.e., assumed to be an ambulatory surgery record): • Length of stay is 0; • Principal procedure is present; • Total charges are nonmissing; • Routine (room and nursing) charges are missing; and • Age in days is not equal to 0. Otherwise, include as an inpatient record.

Connecticut

The HCUP Connecticut files were constructed from files from the Connecticut Health Information Management and Exchange (CHIME), an affiliate of the Connecticut Hospital Association. The files consist of discharge abstract data for inpatient and same-day surgical stays in Connecticut acute care hospitals. Some community hospitals, as defined by the AHA Annual Survey of Hospitals, may not be included in the HCUP Inpatient Databases because their data were not provided by the data source. In 1997, data from 6% of the community hospitals in Connecticut were not received.

Connecticut data are included in HCUP beginning in 1993.

Sample Restrictions. CHIME was to be notified if more than 50% of their hospitals appeared in any year of NIS data. From 1993-1997, the NIS contains less than 50% of the Connecticut hospitals.

Exclusion of Records. The following records were excluded from the HCUP Connecticut data:

- Ambulatory surgery records (records with Patient Type = "A", same-day surgical) were excluded from the HCUP inpatient database.

- Beginning in 1997, discharges with a disposition indicating “patient was admitted as an inpatient to this hospital” were excluded from the HCUP inpatient database. This disposition was not used prior to 1997 and no exclusion was necessary for those years.

Inclusion of Stays in Special Units. Stays in special units within the hospital (e.g., psychiatric, rehabilitation, long-term care) are included in the file.

Shortfall of Discharges in 1995. In 1995, discharges in October are noticeably fewer than in other months by about 25%. This pattern is consistent across all hospitals in the state. No explanation of the shortfall was available from Connecticut Health Information Management and Exchange.

Florida

The HCUP Florida files were constructed from the Florida Hospital Discharge Data Confidential Information received from the Florida Agency for Health Care Administration. The Florida confidential files consist of discharge abstract data from non-federal Florida hospitals. Some community hospitals, as defined by the AHA Annual Survey of Hospitals, may not be included in the HCUP Inpatient Databases because their data were not provided by the data source. In 1997, data from 6% of the community hospitals in Florida were not received.

Florida data are included in HCUP beginning in 1988.

Confidentiality of Records. Florida requested that admission day of week (ADAYWK) be set to missing for all records in the NIS beginning with 1993.

Inclusion of Stays in Special Units. Inpatient stays in special units (e.g., psychiatric, rehabilitation, long-term care) may be included in the HCUP Florida inpatient data. Florida instructs hospitals to submit records only for stays in acute facilities and to exclude records from special units, but according to Florida AHCA, not all hospitals follow these instructions.

Georgia

The HCUP Georgia files were constructed from inpatient files received from GHA - An Association of Hospitals and Health Systems. Inpatient discharge data was provided for hospitals that are a member of GHA. Some community hospitals, as defined by the AHA Annual Survey of Hospitals, may not be included in the HCUP Inpatient Databases because their data were not provided by the data source. In 1997, data from all of the community hospitals in Georgia were received.

Georgia data are included in HCUP beginning in 1997.

Confidentiality of Records. Georgia requested that the race of the patient (RACE) be set to missing for all records in the NIS.

Confidentiality of Physicians. Georgia requested that physician identifiers (MDID_S) be set to missing for all records in the NIS.

Confidentiality of Hospitals. The sample of Georgia hospitals included in the HCUP NIS may not be representative of Georgia hospitals overall because some Georgia hospitals were dropped from the sampling frame to meet confidentiality requirements. Hospitals were dropped from the sampling frame whenever there were fewer than two hospitals in the sampling stratum. This resulted in the exclusion of one hospital from the 1997 sampling frame.

Georgia requested that hospitals not be identified in the NIS database. As a result, the following information was set to missing for all Georgia hospitals:

- Data source hospital identifier (DSHOSPID)
- Hospital state, county FIPS code (HOSPSTCO)
- AHA hospital identifier without leading 6 (IDNUMBER)
- AHA hospital identifier with leading 6 (AHAID)
- Hospital name (HOSPNAME)
- Hospital city (HOSPCITY)
- Hospital address (HOSPADDR), and
- Hospital zip code (HOSPZIP).

The HCUP hospital identifier (HOSPID) can be used to group inpatient records that belong to the same hospital.

In order to further ensure the confidentiality of hospitals, stratifier variables

- Ownership/Control (H_CONTRL),
- Location (H_LOC),
- Teaching status (H_TCH),
- Bedsize (H_BEDSZ), and
- Location, teaching status combined (H_LOCTCH)

were set to missing if the cell defined by H_CONTRL, H_LOC, H_TCH, and H_BEDSZ had fewer than 2 hospitals in the universe of Georgia hospitals. This affected one hospital in 1997.

Exclusion of Records. Records with a discharge disposition of "still a patient" were excluded from the HCUP Georgia data.

Inclusion of Stays in Special Units. The documentation supplied by Georgia does not indicate whether stays in special units within the hospital (e.g., psychiatric, rehabilitation, long-term care) are included in the file.

Hawaii

The HCUP Hawaii files were constructed from inpatient files received from the Hawaii Health Information Corporation (HHIC). Inpatient discharge data was provided for hospitals that are a member of HHIC. Some community hospitals, as defined by the AHA Annual Survey of Hospitals, may not be included in the HCUP Inpatient Databases because their data were not provided by the data source. In 1997, data from 15% of the community hospitals in Hawaii were not received.

Hawaii data are included in the HCUP SID beginning in 1996 and in the HCUP NIS beginning in 1997.

Confidentiality of Hospitals. The sample of Hawaii hospitals included in the HCUP NIS may not be representative of Hawaii hospitals overall because some Hawaii hospitals were dropped from the sampling frame to meet confidentiality requirements. Hospitals were dropped from the sampling frame whenever there were fewer than two hospitals in the sampling stratum. This resulted in the exclusion of six hospitals from the 1997 sampling frame.

Hawaii requested that hospitals not be identified in the NIS database. As a result, the following information was set to missing for all Hawaii hospitals:

- Data source hospital identifier (DSHOSPID)
- Hospital state, county FIPS code (HOSPSTCO)
- AHA hospital identifier without leading 6 (IDNUMBER)
- AHA hospital identifier with leading 6 (AHAID)
- Hospital name (HOSPNAME)
- Hospital city (HOSPCITY)
- Hospital address (HOSPADDR), and
- Hospital zip code (HOSPZIP).

The HCUP hospital identifier (HOSPID) can be used to group inpatient records that belong to the same hospital.

In order to further ensure the confidentiality of hospitals, stratifier variables

- Ownership/Control (H_CONTRL),
- Location (H_LOC),
- Teaching status (H_TCH),
- Bedsize (H_BEDSZ), and
- Location, teaching status combined (H_LOCTCH)

were set to missing if the cell defined by H_CONTRL, H_LOC, H_TCH, and H_BEDSZ had fewer than 2 hospitals in the universe of Hawaii hospitals. This affected no hospitals in 1997.

Exclusion of Records. Records with a discharge disposition of "still a patient" and “admitted as an inpatient to this hospital” were excluded from the HCUP Hawaii data.

Inclusion of Stays in Special Units. The documentation supplied by Hawaii does not indicate whether stays in special units within the hospital (e.g., psychiatric, rehabilitation, long-term care) are included in the file.

Illinois

The HCUP Illinois files were constructed from the Illinois confidential files received from the Illinois Health Care Cost Containment Council (IHCCCC). The Illinois confidential files consist of uniform bills for inpatient stays from Illinois general acute care and specialty hospitals. Some community hospitals, as defined by the AHA Annual Survey of Hospitals, may not be included in the HCUP Inpatient Databases because their data were not provided by the data source. In 1997, data from 0.5% of the community hospitals in Illinois were not received.

Illinois data are included in HCUP beginning in 1988.

Illinois hospitals are required to report 100 percent of discharge records for inpatient stays of at least 24 hours. The IHCCCC reports better than 98 percent compliance with this mandate. If an adjunct skilled nursing facility or nursing home is operated at the same site, these records are not included in the submission to the IHCCCC.

Illinois excludes records with inconsistent data that have not been corrected and records with missing data in IHCCCC-defined required fields from the Illinois source inpatient data.

Sample Restrictions. Illinois requested that no more than 40% of Illinois data appear in any discharge quarter of NIS data.

Confidentiality of Physicians. For 1988-1994, physician identifiers (MDID_S and SURGID_S) for Illinois were set to missing in the NIS data. Beginning in 1995, Illinois does not supply physician identifiers for HCUP.

Inclusion of Stays in Special Units. Stays in skilled nursing facilities or nursing homes attached to a hospital are excluded by Illinois. Stays in other special units within the hospital (e.g., psychiatric, rehabilitation, long-term care) are included in the inpatient discharge data. Stays in specialty hospitals (e.g., children's hospitals, rehabilitation hospitals, etc.) are included in the HCUP Illinois data.

Iowa

The HCUP Iowa files were constructed from the Association of Iowa Hospitals and Health Systems Statewide Database. Iowa supplied discharge abstract data and some uniform bills for acute inpatient discharges from member hospitals. Some community hospitals, as defined by the AHA Annual Survey of Hospitals, may not be included in the HCUP Inpatient Databases because their data were not provided by the data source. In 1997, data from all of the community hospitals in Iowa were received.

Iowa data are included in HCUP beginning in 1988.

Inclusion of Stays in Special Units. The documentation supplied by the data source indicates that the data include stays in acute exempt units, but exclude stays in swing bed and long-term care units.

Kansas

The HCUP Kansas files were constructed from the Kansas Hospital Association inpatient discharge files. These data include inpatient discharge data from general acute care hospitals that are a member of the Kansas Hospital Association. Some community hospitals, as defined by the AHA Annual Survey of Hospitals, may not be included in the HCUP Inpatient Databases because their data were not provided by the data source. In 1997, data from 8% of the community hospitals in Kansas were not received.

Kansas data are included in HCUP beginning in 1993.

Confidentiality of Hospitals. Kansas requested that hospitals not be identified in the NIS database. As a result, the following information was set to missing for all Kansas hospitals:

- Data source hospital identifier (DSHOSPID)
- Hospital state, county FIPS code (HOSPSTCO)
- AHA hospital identifier without leading 6 (IDNUMBER)
- AHA hospital identifier with leading 6 (AHAID)
- Hospital name (HOSPNAME)
- Hospital city (HOSPCITY)
- Hospital address (HOSPADDR), and
- Hospital zip code (HOSPZIP).

The HCUP hospital identifier (HOSPID) can be used to group inpatient records that belong to the same hospital.

Inclusion of Stays in Special Units. The documentation provided by the data source indicates that hospitals are not required to report non-acute discharges, including those from long term care units and facilities. The documentation does not specify whether these discharges and discharges from other special units within a hospital (e.g., psychiatric, rehabilitation, etc.) are excluded from the supplied data.

Maryland

The HCUP Maryland files were constructed from the confidential files received from the State of Maryland's Health Services Cost Review Commission (HSCRC). Demographic and utilization data for inpatient stays in Maryland acute care hospitals were supplied by HSCRC in the Uniform Hospital Discharge Abstract Data Set. Some community hospitals, as defined by the AHA Annual Survey of Hospitals, may not be included in the HCUP Inpatient Databases because their data were not provided by the data source. In 1997, data from all of the community hospitals in Maryland were received.

Maryland data are included in the HCUP SID beginning in 1990 and in the HCUP NIS beginning in 1993.

Inclusion of Stays in Special Units. The documentation provided by the data source does not indicate whether stays in special units within a hospital (e.g., psychiatric, rehabilitation, long-term care) are included in the data.

Massachusetts

The HCUP Massachusetts files were constructed from the Massachusetts confidential Case Mix Database files received from the Massachusetts Division of Health Care Finance and Policy. Massachusetts supplied discharge abstract data for inpatient stays from general acute care hospitals in Massachusetts. Some community hospitals, as defined by the AHA Annual Survey of Hospitals, may not be included in the HCUP Inpatient Databases because their data were not provided by the data source. In 1997, data from 13% of the community hospitals in Massachusetts were not received.

Massachusetts data are included in HCUP beginning in 1988.

Confidentiality of Physicians. All physician identifiers (MDID_S and SURGID_S) for Massachusetts were set to missing in the NIS data starting in 1994.

Inclusion of Stays in Special Units. The documentation provided by the data source indicates that inclusion of discharges from special units within the hospital (e.g., psychiatric, rehabilitation, long-term care) varies by hospital.

Missouri

The HCUP Missouri files were constructed from the Hospital Industry Data Institute (HIDI) inpatient stay files. Missouri supplied discharge abstract data for inpatient stays from Missouri general acute care and specialty hospitals (e.g., children's hospitals, rehabilitation hospitals, and cancer hospitals). Some community hospitals, as defined by the AHA Annual Survey of Hospitals, may not be included in the HCUP Inpatient Databases because their data were not provided by the data source. In 1997, data from 12% of the community hospitals in Missouri were not received.

Missouri data are included in HCUP beginning in 1995.

Sample Restrictions. The sample of Missouri hospitals included in the HCUP NIS may not be representative of Missouri hospitals overall because some Missouri hospitals were dropped from the sampling frame. Hospitals were dropped from the sampling frame if they did not give their permission to be included. This resulted in the exclusion of 35 hospitals from the 1995, 1996, and 1997 sampling frame.

Exclusion of Records. Records with a discharge disposition of "still a patient" were excluded from the HCUP Missouri data.

Inclusion of Stays in Special Units. Missouri supplied discharges from special units within hospitals including psychiatric, rehabilitation, skilled nursing, intermediate care, other long-term care, swing-bed, hospice, and other unspecified inpatient units. Records for these different types of care cannot be identified from data elements included in the HCUP Missouri data.

New Jersey

The HCUP New Jersey files were received from the New Jersey Department of Health and Senior Services. The New Jersey files consist of discharge abstract data for all inpatient and same-day stays. New Jersey supplied discharge abstract data for inpatient stays from general acute care hospitals. Some community hospitals, as defined by the AHA Annual Survey of Hospitals, may not be included in the HCUP Inpatient Databases because their data were not provided by the data source. In 1997, data from 8% of the community hospitals in New Jersey were not received.

New Jersey data are included in HCUP beginning in 1988.

Inclusion of Stays in Special Units. The documentation provided by the data source does not indicate whether stays in special units within the hospital (e.g., psychiatric, rehabilitation, long-term care) are included.

Exclusion of Ambulatory Surgery Records. New Jersey supplied a mixture of inpatient and ambulatory surgery records, which were not distinguished by a record type indicator. Ambulatory surgery records were excluded from the HCUP inpatient database based on a definition supplied by New Jersey. The definition of ambulatory surgery records supplied by New Jersey is:

- Same-day stay (LOS = 0),

- Non-zero charges to operating room or same-day surgery, and

- Discharged to home (DISP = 1).

New York

The HCUP New York files were constructed from the New York State Department of Health's Statewide Planning and Research Cooperative System (SPARCS) Master File. The New York files contain inpatient discharges from acute care hospitals in the state, excluding long-term care units of short-term hospitals and Federal hospitals. Some community hospitals, as defined by the AHA Annual Survey of Hospitals, may not be included in the HCUP Inpatient Databases because their data were not provided by the data source. In 1997, data from 1% of the community hospitals in New York were not received.

New York data are included in the HCUP SID beginning in 1988 and in the HCUP NIS beginning in 1993.

For 1988-1993, New York supplied their Master File which consists of Discharge Data Abstracts (DDAs) matched to Uniform Billing Forms (UBFs) for inpatient stays from all hospitals in the state excluding long-term care units of short-term hospitals and Federal hospitals.

For 1988-1993, New York created the Master File by matching DDAs and UBFs based on Permanent Facility Identifier, Medical Record Number, Admitting Number, Admit Date, and Discharge Date. If the DDA and UBF records matched, the information from the DDA and UBF was included in the Master File. If there was no match, the information from the DDA was included in the Master File. Due to an administrative change in the collection of billing records for 1989, a large percentage of the DDAs could not be matched to a UBF. When there was no match, charge information, which would have come from the UBF, is missing. The match rate improves over time and stabilizes after 1991. The percentage of DDA records that have a matching UBF record in the Master File are as follows:

1988	77.2%
1989	26.3%
1990	62.8%
1991	93.7%
1992	91.8%
1993	95.5%.

Beginning in 1994, hospitals submitted discharge records to New York in a new format, using Universal Data Set (UDS) specifications. This format combines the old UBF and DDA data into a single submission record. In these years, New York supplied records for HCUP that contain complete discharge and uniform billing data corresponding to the "matched" records in earlier years.

Exclusion of Records. The following New York records were excluded from the HCUP inpatient database:

- For all years, interim records for patients who had not been discharged.

- For 1988-1992, records with a transaction code indicating "Deletion of a Record Previously Accepted" were excluded. These records were incorrect versions of accurate records included elsewhere in the SPARCS files. This was not a problem in subsequent years' data.

- For 1988-1993, Uniform Billing Forms (UBFs) that could not be matched to Discharge Data Abstracts (DDAs) were excluded. Matched DDA and UBF records and unmatched DDA records (without charges) were retained in the data.

- Beginning in 1994, records with a discharge disposition of "still a patient."

Inclusion of Stays in Special Units. The documentation supplied by the data source indicates that the data include stays in detoxification (alcohol and drug abuse), alcohol rehabilitation, mental retardation, mental rehabilitation, rehabilitation, alternate level of care, and psychiatric (acute and long term) units within community hospitals. Records for these different types of care cannot be identified from the data elements available in the HCUP New York inpatient data.

Oregon

The 1993-1995 HCUP Oregon files were constructed from the Office for Oregon Health Plan Policy and Research discharge files. Beginning in 1996, HCUP Oregon files were constructed from discharge files supplied by the Oregon Association of Hospitals and Health Systems. The Oregon files consist of discharge abstract data for inpatient stays from member hospitals. Beginning in 1995, discharges from Veteran's Administrations facilities are not reported by the source. Some community hospitals, as defined by the AHA Annual Survey of Hospitals, may not be included in the HCUP Inpatient Databases because their data were not provided by the data source. In 1997, data from 3% of the community hospitals in Oregon were not received.

Oregon data are included in HCUP beginning in 1993.

Exclusion of Records. Beginning in 1995, the source reports the discharge disposition of "still a patient." These records were excluded from the HCUP Oregon data.

Inclusion of Stays in Special Units. Stays in special units within Oregon hospitals (e.g., psychiatric, rehabilitation, long-term care) are included in the source data and therefore in the HCUP inpatient database.

Pennsylvania

The HCUP Pennsylvania files were constructed from the Pennsylvania Health Care Cost Containment Council files. Pennsylvania supplied uniform bills from general acute care, state psychiatric, and rehabilitation facilities and from children's and specialty hospitals. Some community hospitals, as defined by the AHA Annual Survey of Hospitals, may not be included in the HCUP Inpatient Databases because their data were not provided by the data source. In 1997, data from 3% of the community hospitals in Pennsylvania were not received.

Pennsylvania data are included in HCUP beginning in 1989.

Confidentiality of Records. Pennsylvania requested that patient age (AGE and AGEDAY) be set to the midpoint of 5-year intervals for records in the NIS with the following sensitive conditions: abortion, AIDS, mental illness, and substance abuse. See Pennsylvania note under the data elements AGE and AGEDAY for information on how these conditions were defined.

Exclusion of Records. Records with a discharge disposition of "still a patient" were excluded from the HCUP Pennsylvania data.

Inclusion of Stays in Special Units. Pennsylvania supplied discharges from psychiatric, drug and alcohol, and rehabilitation units of general acute care hospitals. Records for these different types of care cannot be identified from data elements included in the HCUP Pennsylvania data.

South Carolina

The HCUP South Carolina files were constructed from confidential data files supplied by the South Carolina State Budget and Control Board. The data include inpatient stays from South Carolina acute care hospitals. Some community hospitals, as defined by the AHA Annual Survey of Hospitals, may not be included in the HCUP Inpatient Databases because their data were not provided by the data source. In 1997, data from 8% of the community hospitals in South Carolina were not received.

South Carolina data are included in HCUP beginning in 1993.

Confidentiality of Hospitals. The sample of South Carolina hospitals included in the HCUP NIS may not be representative of South Carolina hospitals overall because some South Carolina hospitals were dropped from the sampling frame to meet confidentiality requirements. Hospitals were dropped from the sampling frame whenever there were fewer than two hospitals in the sampling stratum. This resulted in the exclusion of:

- five hospitals from 1993,
- four hospitals from 1994,
- four hospitals from 1995,
- six hospitals from 1996, and
- six hospitals from the 1997 sampling frame.

South Carolina requested that hospitals not be identified in the NIS database. As a result, the following information was set to missing for all South Carolina hospitals:

- Data source hospital identifier (DSHOSPID)
- Hospital state, county FIPS code (HOSPSTCO)
- AHA hospital identifier without leading 6 (IDNUMBER)
- AHA hospital identifier with leading 6 (AHAID)
- Hospital name (HOSPNAME)
- Hospital city (HOSPCITY)
- Hospital address (HOSPADDR), and
- Hospital zip code (HOSPZIP).

The HCUP hospital identifier (HOSPID) can be used to group inpatient records that belong to the same hospital.

In order to further ensure the confidentiality of hospitals, stratifier variables

- Ownership/Control (H_CONTRL),
- Location (H_LOC),
- Teaching status (H_TCH),
- Bedsize (H_BEDSZ), and

- Location, teaching status combined (H_LOCTCH)

were set to missing if the cell defined by H_CONTRL, H_LOC, H_TCH, and H_BEDSZ had fewer than 2 hospitals in the universe of South Carolina hospitals. This affected three hospitals in 1993, and one hospital in 1994-1997.

Exclusion of Records. The following records were excluded from the HCUP South Carolina data:

- Beginning in 1994, discharges with disposition of "still a patient" were excluded from the HCUP inpatient database. This disposition was not used in 1993 and no exclusion was necessary for that year.

- Beginning in 1996, discharges with a disposition indicating “patient was admitted as an inpatient to this hospital” were excluded from the HCUP inpatient database. This disposition was not used prior to 1997, and no exclusion was necessary for those years.

Inclusion of Stays in Special Units. The documentation supplied by South Carolina indicates that stays in long term care units and facilities were excluded by South Carolina from the supplied data.

Tennessee

The HCUP Tennessee files were constructed from the inpatient files received from THA - An Association of Hospitals and Health Systems. These data include inpatient discharge data from Tennessee general acute care and some specialty facilities (e.g., children's hospitals, rehabilitation hospitals, state psychiatric facilities, etc.) that are members of THA. Some community hospitals, as defined by the AHA Annual Survey of Hospitals, may not be included in the HCUP Inpatient Databases because their data were not provided by the data source. In 1997, data from 22% of the community hospitals in Tennessee were not received.

Tennessee data are included in HCUP beginning in 1995.

Confidentiality of Hospitals. The sample of Tennessee hospitals included in the HCUP NIS may not be representative of Tennessee hospitals overall because some Tennessee hospitals were dropped from the sampling frame to meet confidentiality requirements. Hospitals were dropped from the sampling frame whenever there were fewer than two hospitals in the sampling stratum. This resulted in the exclusion of:

- six hospitals from 1995,
- four hospitals from 1996, and
- five hospitals from the 1997 sampling frame.

Tennessee requested that hospitals not be identified in the NIS database. As a result, the following information was set to missing for all Tennessee hospitals:

- Data source hospital identifier (DSHOSPID)
- Hospital state, county FIPS code (HOSPSTCO)
- AHA hospital identifier without leading 6 (IDNUMBER)
- AHA hospital identifier with leading 6 (AHAID)
- Hospital name (HOSPNAME)
- Hospital city (HOSPCITY)

- Hospital address (HOSPADDR), and
- Hospital zip code (HOSPZIP).

The HCUP hospital identifier (HOSPID) can be used to group inpatient records that belong to the same hospital.

In order to further ensure the confidentiality of hospitals, stratifier variables

- Ownership/Control (H_CONTRL),
- Location (H_LOC),
- Teaching status (H_TCH),
- Bedsize (H_BEDSZ), and
- Location, teaching status combined (H_LOCTCH)

were set to missing if the cell defined by H_CONTRL, H_LOC, H_TCH, and H_BEDSZ had fewer than 2 hospitals in the universe of Tennessee hospitals. This affected no hospitals in 1995-1997.

Exclusion of Records. The following records were excluded from the HCUP Tennessee data:

- Records with a discharge disposition of "still a patient."

- Continuation records that only contained information on additional detailed charges.

- Beginning in 1996, discharges with a disposition indicating “patient was admitted as an inpatient to this hospital” were excluded from the HCUP inpatient database. Due to an error in HCUP processing, these records were retained in the 1995 HCUP Tennessee inpatient data. These affected discharges in 1995 can be identified by the discharge disposition of invalid (DISP = .A).

Inclusion of Stays in Special Units. The documentation supplied by Tennessee indicates that stays in special units within the hospital (e.g., psychiatric, rehabilitation, long-term care) are included in the file.

Utah

The HCUP Utah files were constructed from inpatient files received from Office of Health Data Analysis, Utah Department of Health. These data include inpatient discharge data from Utah general acute care and some specialty facilities (e.g., children's hospitals, rehabilitation hospitals, state psychiatric facilities, etc.) associated with acute care hospitals. Some community hospitals, as defined by the AHA Annual Survey of Hospitals, may not be included in the HCUP Inpatient Databases because their data were not provided by the data source. In 1997, data from 2% of the community hospitals in Utah were not received.

Utah data are included in HCUP beginning in 1997.

Confidentiality of Physicians. Utah requested that physician identifiers (MDID_S and SURGID_S) be set to missing for all records in the NIS.

Inclusion of Stays in Special Units. The documentation supplied by Utah does not indicate whether stays in special units within the hospital (e.g., psychiatric, rehabilitation, long-term care) are included in the file.

Washington

The HCUP Washington files were constructed from the Washington Comprehensive Hospital Abstract Reporting System (CHARS) data received from the Washington State Department of Health. Washington supplied uniform bills for inpatient stays from all acute care units, alcohol dependency units, bone marrow transplant units, extended care units, psychiatric units, rehabilitation units, group health units, and swing bed units. Some community hospitals, as defined by the AHA Annual Survey of Hospitals, may not be included in the HCUP Inpatient Databases because their data were not provided by the data source. In 1997, data from 1% of the community hospitals in Washington were not received.

Washington data are included in HCUP beginning in 1988.

Inclusion of Stays in Special Units. The documentation provided by the data source indicates that stays in special units within a hospital are included in the data. Records for these different types of care can be identified by the fourth digit of the source-supplied hospital identifier (DSHOSPID) on each patient record:

None	=	General acute care
A	=	Alcohol Dependency Unit
B	=	Bone Marrow Transplant Unit
E	=	Extended Care Unit
H	=	Tacoma General/Group Health Combined
I	=	Group Health only at Tacoma Hospital
P	=	Psychiatric Unit
R	=	Rehabilitation Unit
S	=	Swing Bed Unit

Washington assigns this value to DSHOSPID based upon the type of unit discharging the patient.

Wisconsin

The HCUP Wisconsin files were constructed from confidential files received from the Bureau of Health Information, Wisconsin Department of Health and Family Services. Wisconsin supplied discharge data abstract and uniform bills for non-federal Wisconsin hospitals. Some community hospitals, as defined by the AHA Annual Survey of Hospitals, may not be included in the HCUP Inpatient Databases because their data were not provided by the data source. In 1997, data from all of the community hospitals in Wisconsin were received.

Wisconsin data are included in HCUP beginning in 1989.

Inclusion of Stays in Special Units. The documentation supplied by the data source does not indicate whether stays in special units within a hospital (e.g., psychiatric, rehabilitation, long-term care) are included in the data.

TECHNICAL SUPPLEMENT 13:
HCUP DATA QUALITY TABLE

This Technical Supplement provides information on the results of edit checks performed on the Nationwide Inpatient Sample. For details on these edit checks, refer to Technical Supplement: *Quality Control in HCUP Data Processing*.

HCUP DATA QUALITY TABLE

HCUP DATA QUALITY TABLE
NIS 1989-1997 INPATIENT DATA

Description	Edit Checks	1989	1990	1991	1992	1993	1994	1995	1996	1997
Number of Hospitals and Discharges										
Total Number of Hospitals		882 (100.00%)	871 (100.00%)	859 (100.00%)	856 (100.00%)	913 (100.00%)	904 (100.00%)	938 (100.00%)	906 (100.00%)	1,012 (100.00%)
Total Discharges		6,110,064 (100.00%)	6,268,515 (100.00%)	6,156,188 (100.00%)	6,195,744 (100.00%)	6,538,976 (100.00%)	6,385,011 (100.00%)	6,714,935 (100.00%)	6,542,069 (100.00%)	7,148,420 (100.00%)
Overall Data Quality Indicators										
Failed any edit or validity check	ED010-ED952, DXV1-DXV30=1 or PRV1-PRV25=1	227,284 (3.72%)	140,895 (2.25%)	151,976 (2.47%)	285,492 (4.61%)	184,701 (2.82%)	267,238 (4.19%)	349,457 (5.20%)	327,577 (5.01%)	370,673 (5.19%)
Failed any edit check	ED010-ED952	185,412 (3.03%)	135,766 (2.17%)	135,923 (2.21%)	275,382 (4.44%)	141,776 (2.17%)	241,805 (3.79%)	340,702 (5.07%)	318,095 (4.86%)	352,071 (4.93%)
Edit Checks										
Reported LOS not equal to calculated LOS	ED010	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	2,585 (0.04%)	0 (0.00%)	0 (0.00%)	1 (0.00%)	203 (0.00%)

HCUP DATA QUALITY TABLE
NIS 1989-1997 INPATIENT DATA

Description	Edit Checks	1989	1990	1991	1992	1993	1994	1995	1996	1997
Admission date after discharge date	ED011	13 (0.00%)	22 (0.00%)	23 (0.00%)	20 (0.00%)	78 (0.00%)	152 (0.00%)	74 (0.00%)	14 (0.00%)	21 (0.00%)
Reported AGE not equal to calculated AGE	ED020	13,625 (0.22%)	12,622 (0.20%)	12,618 (0.20%)	13,484 (0.22%)	8,110 (0.12%)	663 (0.01%)	4,350 (0.06%)	964 (0.01%)	1,375 (0.02%)
Age in years inconsistent with AGEDAY	ED021	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)	18 (0.00%)	0 (0.00%)	0 (0.00%)
Long length of stay (LOS), justified	ED600	463 (0.01%)	370 (0.01%)	362 (0.01%)	331 (0.01%)	368 (0.01%)	294 (0.00%)	407 (0.01%)	358 (0.01%)	383 (0.01%)
Long length of stay (LOS), unjustified	ED601	55 (0.00%)	45 (0.00%)	60 (0.00%)	65 (0.00%)	80 (0.00%)	81 (0.00%)	82 (0.00%)	46 (0.00%)	43 (0.00%)
Low charges/day, justified	ED910	2,378 (0.04%)	1,014 (0.02%)	813 (0.01%)	1,606 (0.03%)	806 (0.01%)	2,405 (0.04%)	1,698 (0.03%)	837 (0.01%)	1,431 (0.02%)
Low charges/day, unjustified	ED911	14,979 (0.25%)	5,952 (0.09%)	4,497 (0.07%)	7,944 (0.13%)	4,370 (0.07%)	10,237 (0.16%)	8,510 (0.13%)	6,440 (0.10%)	8,700 (0.12%)
High charges/day, justified	ED920	846 (0.01%)	1,283 (0.02%)	1,620 (0.03%)	2,393 (0.04%)	2,560 (0.04%)	3,105 (0.05%)	3,675 (0.05%)	3,681 (0.06%)	4,803 (0.07%)
High charges/day, unjustified	ED921	253 (0.00%)	594 (0.01%)	815 (0.01%)	1,281 (0.02%)	2,102 (0.03%)	3,694 (0.06%)	9,716 (0.14%)	10,614 (0.16%)	16,541 (0.23%)

HCUP DATA QUALITY TABLE
NIS 1989-1997 INPATIENT DATA

Description	Edit Checks	1989	1990	1991	1992	1993	1994	1995	1996	1997
Unacceptable uniform payer combination	ED951	114,540 (1.87%)	114,019 (1.82%)	113,310 (1.84%)	121,921 (1.97%)	112,156 (1.72%)	123,998 (1.94%)	268,880 (4.00%)	290,946 (4.45%)	315,741 (4.42%)
Unacceptable non-unif payer combination	ED952	147 (0.00%)	246 (0.00%)	520 (0.01%)	304 (0.00%)	317 (0.00%)	213 (0.00%)	799 (0.01%)	1,106 (0.02%)	748 (0.01%)
Edit Checks on Diagnoses (DX) and Procedures (PR)										
Principal DX inconsistent w/ age, sex	ED101, ED301, ED401	447 (0.01%)	403 (0.01%)	321 (0.01%)	268 (0.00%)	653 (0.01%)	361 (0.01%)	263 (0.00%)	198 (0.00%)	277 (0.00%)
Secondary DX inconsistent w/ age, sex	ED102-ED130, ED302-ED330, ED402-ED430	2,486 (0.04%)	1,554 (0.02%)	1,522 (0.02%)	1,526 (0.02%)	2,042 (0.03%)	2,005 (0.03%)	1,837 (0.03%)	1,589 (0.02%)	1,929 (0.03%)
Principal PR inconsistent w/ age, sex	ED201, ED501	341 (0.01%)	352 (0.01%)	251 (0.00%)	391 (0.01%)	548 (0.01%)	268 (0.00%)	225 (0.00%)	213 (0.00%)	211 (0.00%)
Secondary PR inconsistent w/ age, sex	ED202-ED225, ED502-ED525	147 (0.00%)	177 (0.00%)	137 (0.00%)	136 (0.00%)	257 (0.00%)	230 (0.00%)	125 (0.00%)	170 (0.00%)	247 (0.00%)
Day for principal PR w/o procedure coded	ED701	72 (0.00%)	12 (0.00%)	1,247 (0.02%)	14 (0.00%)	424 (0.01%)	93,819 (1.47%)	49,060 (0.73%)	1,267 (0.02%)	50 (0.00%)

HCUP DATA QUALITY TABLE
NIS 1989-1997 INPATIENT DATA

Description	Edit Checks	1989	1990	1991	1992	1993	1994	1995	1996	1997
Day for principal PR not during stay	ED801	38,760 (0.63%)	593 (0.01%)	1,135 (0.02%)	127,750 (2.06%)	6,481 (0.10%)	1,036 (0.02%)	4,508 (0.07%)	820 (0.01%)	716 (0.01%)
Validity Checks on Diagnoses (DX) and Procedures (PR)										
Principal diagnosis invalid	DXV1=1	32,369 (0.53%)	408 (0.01%)	2,847 (0.05%)	2,618 (0.04%)	13,059 (0.20%)	5,136 (0.08%)	1,272 (0.02%)	1,340 (0.02%)	2,798 (0.04%)
Secondary diagnosis invalid	DXV2-DXV25=1	9,125 (0.15%)	1,739 (0.03%)	6,902 (0.11%)	4,705 (0.08%)	29,868 (0.46%)	18,054 (0.28%)	3,902 (0.06%)	2,753 (0.04%)	14,536 (0.20%)
Principal procedure invalid	PRV1=1	4,688 (0.08%)	491 (0.01%)	2,810 (0.05%)	1,699 (0.03%)	1,865 (0.03%)	438 (0.01%)	1,273 (0.02%)	2,032 (0.03%)	817 (0.01%)
Secondary procedure invalid	PRV2-PRV25=1	1,404 (0.02%)	2,640 (0.04%)	3,914 (0.06%)	1,524 (0.02%)	2,279 (0.03%)	2,968 (0.05%)	3,283 (0.05%)	3,888 (0.06%)	2,431 (0.03%)

Note: This Technical Supplement is based on 1995 data—an updated analysis on 1997 data will be available in Spring 2000.

TECHNICAL SUPPLEMENT 14: COMPARATIVE ANALYSIS OF HCUP AND NHDS INPATIENT DISCHARGE DATA

EXECUTIVE SUMMARY

This report assesses potential biases of statistics calculated from the Nationwide Inpatient Sample (NIS), Release 4 of the Healthcare Cost and Utilization Project (HCUP). The NIS, Release 4 includes hospital discharge data from a sample of community hospitals for calendar year 1995. Statistics for discharge- and hospital-level characteristics of the NIS data are compared with the National Hospital Discharge Survey (NHDS) data.

Most statistics calculated from the NIS are consistent with those from the NHDS, particularly those for region and patient characteristics. Several differences exist between the NIS and NHDS discharge estimates when discharges are stratified by hospital size. The sample of hospitals in the NIS was stratified on hospital size and weighted to the AHA universe to better represent the universe of hospitals. The NIS estimates of average length of stay appear consistent with the NHDS. NIS estimates of in-hospital mortality rates are higher than the NHDS estimates in all the regions except the Northeast.

Inconsistencies between the NIS estimates and estimates from the NHDS data may be caused by a number of factors. Sample design may cause some differences. Some may be due to differences in coding schemes. In other cases, differences may be attributed to slightly dissimilar populations.

INTRODUCTION

This report assesses potential biases of statistics calculated from the Nationwide Inpatient Sample (NIS), Release 4 of the Healthcare Cost and Utilization Project (HCUP). The NIS, Release 4 includes hospital discharge data from a sample of community hospitals for the calendar year 1995. Statistics for discharge- and hospital-level characteristics of the NIS data are compared with the National Hospital Discharge Survey (NHDS) and the American Hospital Association Annual Survey data.

The NIS, Release 4 was established to provide analyses of hospital utilization across the United States. For each calendar year, the NIS *universe* of hospitals was established as all community hospitals located in the U.S. However, the NIS *sampling frame* was constructed from the subset of universe hospitals that released their discharge data for research use. Currently, the Agency for Health Care Policy and Research (AHCPR) has agreements with 22 data sources that maintain statewide, all-payer discharge data files to include their data in the HCUP database. However, only 19 of these states could be included for this fourth release. These 19 states represent the addition of two states more than the second and third releases, and eight states more than the first release, as shown by Table 56. The NIS, Release 4 is composed of all discharges from a sample of hospitals from these frame states.

Table 56. States in the Frame for the NIS, Release 4

Calendar Years	States in the Frame
1988 (Release 1)	California, Colorado, Florida, Illinois, Iowa, Massachusetts, New Jersey, and Washington
1989-1992 (Release 1)	Add Arizona, Pennsylvania, and Wisconsin
1993 (Release 2) 1994 (Release 3)	Add Connecticut, Kansas, Maryland, New York, Oregon, and South Carolina
1995 (Release 4)	Add Missouri and Tennessee

Creation of the NIS was subject to certain restrictions.

- The Illinois Health Care Cost Containment Council stipulated that no more than 40 percent of Illinois discharge data could be included in the database for any calendar quarter. Consequently, approximately 50 percent of the Illinois community hospital universe was randomly selected for the frame each year.

- Hospitals in Missouri were allowed to withhold their data from the NIS. Thirty-five Missouri hospitals, from a state total of 119, chose not to participate in the NIS.

- South Carolina and Tennessee both imposed "small strata/cell restrictions," requiring the NIS to exclude hospitals, when only one state hospital appears in a sampling strata. As a result, the NIS is not representative of South Carolina or Tennessee hospitals.

To improve the generalizability of the NIS estimates, five hospital sampling strata were used:

1. *Geographic Region* — Midwest, Northeast, West, and South.

2. *Ownership* — government, investor-owned, and nonprofit nongovernment.

3. *Location* — urban and rural.

4. *Teaching Status* — teaching and non-teaching.

5. *Bedsize* — small, medium, and large, specific to the hospital's location and teaching status as shown in Table 57.

Table 57. Bedsize Categories

Location and Teaching Status	Bedsize		
	Small	Medium	Large
Rural	1-49	50-99	100+
Urban, non-teaching	1-99	100-199	200+
Urban, teaching	1-299	300-499	500+

To further ensure geographic representativeness, hospitals were sorted by state and the first three digits of their zip code prior to systematic sampling.

The NIS is a stratified probability sample of hospitals in the frame, with sampling probabilities calculated to select 20 percent of the universe contained in each stratum. The overall objective was to select a sample of hospitals "generalizable" to the target universe, including hospitals outside the frame (which had a zero probability of selection). See *Design of the HCUP Nationwide Inpatient Sample, Release 4,* for more details on the design of the sample.

Sample weights were developed for the NIS to obtain national estimates of hospital and inpatient parameters. For example, with these weights it should be possible to estimate DRG-specific average lengths of stay over all U.S. hospitals, using weighted average lengths of stay based on averages or regression estimates from the NIS. Ideally, relationships among outcomes and their correlates estimated from the NIS should generally hold across all U.S. hospitals. However, since only 19 states contributed data to this fourth release, some estimates may be biased. In this report, we compare estimates based solely on the NIS against estimated quantities from other sources of data.

This report compares both discharge- and hospital-level statistics. Discharge statistics include discharge counts, inpatient charges, in-hospital mortality, and average lengths of stay. Hospital statistics include items such as number of beds, occupancy rates, and staffing levels.

This report is divided into four sections. The first section includes a discussion of the data sources used in the analysis. The second section explains the methodology used to compare the NIS and NHDS. The third section includes a presentation of the results: tables for this section are included at the end of the report. The final section offers some conclusions and recommendations for analyses of the NIS, Release 4.

DATA SOURCES

Benchmark statistics for 1995 from several data sources were compared. The NIS, Release 4, 1995 data were drawn from a frame of 19 states and includes approximately 6.7 million discharges from 938 hospitals. NIS statistics were mainly compared with those calculated from these two data sources:

1. *National Hospital Discharge Survey (NHDS), 1995*. Conducted by the National Center for Health Statistics, the NHDS includes about 260,000 discharges sampled from 400 hospitals. To be part of the NHDS, hospitals must have six or more beds staffed for patient use. The NHDS covers discharges from short-stay U.S. hospitals (hospitals with an average length of stay under 30 days), general-specialty (medical or surgical) hospitals, and children's hospitals. Federal, military, and Veterans Administration hospitals are excluded from the survey. The NHDS sampling frame includes very few specialty hospitals such as psychiatric, maternity, alcohol/chemical dependency, orthopedic, and head-injury hospitals.

   Statistics calculated from the NHDS do have sampling error. However, the statistics are assumed to be unbiased because the sampling frame is relatively unrestricted, encompassing all nonfederal, acute-care, general U.S. hospitals with six or more beds.

2. *AHA Annual Survey of Hospitals, 1995*. This hospital-level file contains one record for every hospital in the NIS universe, making it a convenient source for calculating various statistics based on both the population of hospitals and the NIS sample of hospitals. The file contains hospital-level statistics for hospital reporting periods, which do not necessarily correspond to the calendar year.

Table 58 summarizes some of the key differences in hospitals and discharges represented by the NIS and NHDS data files.

METHODS

Comparisons with NHDS

The following measures were chosen to compare the NIS and NHDS databases:

- Total number of discharges

- Average length of stay (ALOS)

- In-hospital mortality rate

These measures of utilization and outcomes were selected because they are typically used in health services research.

For each statistic, a test was performed to determine whether a difference was statistically significant between the NIS and NHDS estimates. Since the NHDS estimate was based on a sample, two-sample t-tests were used, as described in the Appendix. Differences were reported at the one and five percent significance levels.

To assess their reliability, the statistics listed above were compared within the following types of strata:

- Geographic regions (Midwest, Northeast, West, and South)

- Hospital characteristics (ownership, rural location, teaching status, and bedsize)

- Patient characteristics (age, race, gender, and payer)

- Diagnosis groups (The principal diagnosis code for each discharge was assigned to a diagnosis group defined by the Clinical Classifications for Health Policy Research (CCHPR), Version 2 algorithm — see Elixhauser and McCarthy, 1996).

- Procedure groups (The principal procedure code for each discharge was assigned to a procedure group defined by the CCHPR, Version 2 algorithm — see Elixhauser and McCarthy, 1996).

Further, special analyses were conducted for hospitals in the South region, an area in which the NIS coverage is limited. In the NIS, Release 1, the South region was represented by only Florida. The Second Release of the NIS added Maryland and South Carolina. For Release 4 of the NIS, the South is represented by Florida, Maryland, South Carolina, and Tennessee.

All NIS statistics used sample weights and accounted for the sample design using the SUDAAN microcomputer statistical software to calculate finite sample statistics and their variances. All NHDS statistics were calculated with Statistical Analysis System (SAS) microcomputer software. For NHDS statistics, standard errors were calculated as described in the Appendix.

RESULTS

Comparisons Between the NIS and the NHDS

Since the NIS and the NHDS represent different samples of the same universe of hospitals, some differences are expected, and can be attributed to statistical "noise." Moreover, because of the large number of comparisons, some of the statistically significant differences will not be real differences using 0.05 level of significance. While bias could be present in either sample, the NHDS estimates are less likely to be biased because the hospital sampling frame is far less restricted than that for the NIS. The following sections describe results of statistical comparisons by region, hospital characteristics, patient characteristics, diagnosis, and procedure.

Comparisons by Region

Table 59 compares estimates of discharges, average lengths of stay, and in-hospital mortality generated from NIS and NHDS data. Comparisons are presented by total and by region for 1995. The NIS and NHDS estimates of national and regional discharges do not significantly differ. Overall, the NIS and NHDS produce similar estimates of average length of stay, although the NIS estimate is significantly higher than the NHDS estimate for the Midwest (by 30 percent). NIS in-hospital mortality rate estimates are also significantly higher in total (by 8 percent) for the Midwest and South (by 24 and 12 percent respectively).

Comparisons by Hospital Characteristics

Table 60 compares estimates of discharges, average lengths of stay, and in-hospital mortality between the NIS and NHDS for 1995, by hospital ownership categories (private/investor-owned, private/nonprofit, and government/nonfederal) and bedsize categories (6-99, 100-199, 200-299, 300-499, and 500+).

Several of the estimates for hospital discharges differ significantly between the two sources. For government hospitals, the NIS estimates 15 percent more discharges than the NHDS. For private hospitals, which represent the majority of the discharges, there is no significant difference in total discharges for either nonprofit or investor-owned hospitals. Within the ownership groups, significant differences are found for most bedsize categories except for 200-299 bed hospitals. The NIS estimates more discharges than the NHDS for five of the 10 significant differences, and fewer for the remaining five.

It should be noted that the total number of 1995 universe discharges in hospitals with over 500 beds is 6.6 million according to the AHA file. Consequently, the NIS (with 7.0 million) may provide a better estimate of discharge counts for large hospitals than the NHDS (with 3.9 million). These differences in estimated discharge counts may contribute to differences in outcome statistics, reported in Table 60, between the two sources because the discharge counts are essentially sums of discharge weights, which are used to calculate outcome statistics.

Totals for each ownership group show no significant differences in average length of stay (ALOS) or in-hospital mortality estimates. In addition, there are few differences within the ownership groups between the two sources: we note here one significant ALOS difference out of 15 comparisons. A significant ALOS difference between the NIS and NHDS for government hospitals is found only for 100-199 bed hospitals (19 percent higher).

Estimates for in-hospital mortality tend to be higher for the NIS than for NHDS, although not significantly in most cases. There are only four significant differences between the NIS and NHDS estimates although the NIS estimate is higher than the NHDS estimate for 12 of the 15 strata. The NIS estimate is significantly higher than the NHDS estimate for investor-owned hospitals with 100-199 beds (by 15 percent), and for nonprofit hospitals with fewer than 6-99 beds (31 percent) and between 100-199 beds (by 16 percent).

Comparisons by Patient Characteristics

Table 61 compares estimates of discharges, average lengths of stay, and in-hospital mortality between the NIS and NHDS for 1995 — by primary payer, age group, gender, and race. The NIS contains uniform values for race, however, there is variation in source data from the participating states. Specifically, in some states hospitals report “other” race for all non-white patients, resulting in overreporting for this race category. Any analysis of NIS data by race categories is affected by this variation. Except for mortality, the majority of estimates are not significantly different between the two data sources for these strata.

Discharge estimates for Medicare, Medicaid, private insurance, all age groups, males, females, and three categories of race (White, Black, and missing), show no significant differences between the NIS and NHDS. Significant differences however, are found for the payer categories of self-pay, no charge, other, and missing. The NIS discharge estimates for self-pay patients is 40 percent higher than the NHDS estimate. For no charge, other, and missing payer, the NIS estimates are lower than the NHDS estimates. The NIS estimate for other race is higher than the NHDS estimate by 158 percent.

Average length of stay estimates from the two sources are not statistically different. Estimates of in-hospital mortality rates from the NIS also tend to be higher than the NHDS estimates. Of the 17 strata, the NIS estimates are larger than the NHDS estimates for 11 strata, although not all differences are statistically significant. The NIS estimates are significantly larger than NHDS estimates for the payer category of other (36 percent); age groups 15-44 years, and 65+ years (17 and 4 percent); males and females (6 and 9 percent); plus the white, and missing race categories (12 and 18 percent). The NIS estimate is significantly smaller, by 16 and 24 percent respectively, than the NHDS estimate for the age group 0-15 years and other race strata.

Comparisons for the South Region

Table 62 gives a detailed comparison for the South Region by hospital and patient characteristics. Of the 21 strata in Table 62, significant differences are found between the NIS and NHDS estimates for discharges (8 out of 21) and in-hospital mortality rates (6 out of 21). None of the comparisons for average lengths of stay are statistically different.

No significant differences in discharge estimates are found for any ownership, age group, or gender category. Four of the five bedsize categories, however, show significant differences between the NIS and NHDS estimates of discharges. The NIS estimates are lower than the NHDS estimates for small and medium hospitals (6-99, 100-199, and 200-299 beds) by 9 to 28 percent. The NIS estimates for very large hospitals (500+ beds) are larger than the NHDS estimates by 53 percent. No significant differences are found for the primary payer categories of Medicare, Medicaid, and private insurance, while the categories of self-pay, no charge, other and missing do show significant differences. NIS discharge estimates are higher for the self-pay category and lower for the no charge, other, and missing categories. These are similar to the discharge estimates over all regions by payer as found in Table 61.

The average length of stay estimates from the NIS generally agree with the NHDS estimates for the South. The NIS in-hospital mortality estimates are higher than the NHDS estimates for nearly every hospital and patient category, including by age group (17 of the 23 strata), although only six of the differences are significant. The higher NIS estimates may stem from the large impact of Florida hospitals on the estimate for the South. Florida accounts for 52% of Southern discharges and 51% of Southern hospitals within the 1995 NIS data. Because many of the Southern states are not represented in the NIS, discharges from Florida hospitals, and the characteristics of Florida's hospital and patient populations, may be amplified in NIS estimates. Specifically, Florida has a large immigrant population with serious health problems and this may explain some of the differences in mortality estimates.

Comparisons by Diagnosis Category

Table 63 compares the NIS and NHDS by the 25 most frequent primary diagnosis categories, ranked according to the NIS estimates of number of discharges for each category. CCHPR code categories (version 2) are assigned based on the primary (vs. principal or admitting) diagnosis. The NIS discharge estimates differ significantly from the NHDS estimates for 12 of the 25 CCHPR categories; NIS estimates are significantly higher for eight diagnosis categories and significantly lower for four categories.

Some of the discrepancies found in the estimated number of discharges may be explained by differences in the assignment of primary diagnosis for the NIS and NHDS databases. In building the NIS, there is no reordering of diagnoses. The first diagnosis listed for each discharge was assigned as the primary diagnosis (although the state organizations that supply NIS data may have assigned the principal diagnoses to the primary diagnosis position prior to supplying data for the NIS). The NHDS reordered diagnoses under certain conditions.

For example, differences in the number of delivery-related discharges could be explained by the reordering of diagnosis codes in the NHDS. For women discharged after a delivery, a code of V27 (Outcome of Delivery) from the supplemental classification is entered as the second-listed code. A code designating normal or abnormal delivery is then listed in the first position. This could explain differences in the number of discharges counted in the diagnosis group for normal pregnancy and/or delivery (rank 8), trauma to the perineum and vulva (rank 6), fetal distress and abnormal forces of labor (rank 18), other complications of birth affecting mother (rank 23), and other complications of pregnancy (rank 24).

As another example of diagnosis reordering in the NHDS, if the first-listed diagnosis was a symptom, it was reassigned as a secondary diagnosis. This may have affected estimates for the 13th ranked diagnosis category, nonspecific chest pain. Taking into account the differences in ordering of diagnoses reduces the number of significant differences in estimated discharges between the two data sources from 12 to six of the 25 categories.

Comparisons of ALOS and in-hospital mortality rates by diagnosis category (also shown in Table 63) indicate few significant differences between NIS and NHDS estimates. Significant differences are found for only one ALOS estimate (Normal Pregnancy) and for no in-hospital mortality estimates. The in-hospital mortality rates yielded valid significance tests for only 19 categories. This is due to the fact that valid NHDS standard errors for in-hospital mortality could not be calculated for six categories (see Appendix for validity criteria).

Comparisons by Procedure Category

Table 64 lists the top 25 procedure categories, ranked according to the NIS estimates of number of discharges for each category. Similar to the diagnosis groups, CCHPR codes are assigned based on the primary, or first-listed, procedure for each discharge. The NIS discharge estimates differ significantly from the NHDS estimates for nine of the 25 CCHPR categories; NIS estimates are significantly higher for seven procedure categories, and significantly lower for only two categories.

Procedures for which the NIS discharges were significantly higher than the NHDS estimates include the following: episiotomy, diagnostic cardiac catheterization, upper GI, percutaneous coronary angioplasty, respiratory intubation, CT head scans, and cancer chemotherapy. These differences may be explained by the estimated high number of discharges from large hospitals in the NIS, which are more likely to perform high technology procedures (see Table 60), compared to the number of large hospitals in NHDS.

Comparisons of average length of stay and in-hospital mortality rate estimates by procedure category show few significant differences between NIS and NHDS estimates. Three significant differences are found for ALOS, and three differences are also found for in-hospital mortality. Significance tests were not performed for five in-hospital mortality rate estimates due to the unavailability of valid standard errors for NHDS estimates (see Appendix).

Comparison with AHA Data

Table 65 demonstrates that hospital weights associated with the NIS yield hospital counts consistent with AHA universe counts for various categories of hospital types. This is expected because the sample of NIS hospitals was stratified on most of these variables, and sample hospital weights were calculated within strata based on AHA data.

Table 66 compares the universe (AHA) and weighted frame (NIS) means and medians for selected hospital-level measures defined in the 1995 AHA Annual Survey. In general, the frame hospital weighted averages and medians tend to be slightly higher than the universe averages.

DISCUSSION

In general, for many types of estimates, the NIS performs very well. Some differences emerge when the NIS is compared to specific data sets. Sometimes, these variations are caused by differences in definitions (e.g., NIS and NHDS coding schemes). In some cases, differences are due to certain shortcomings in the NIS.

Comparisons of Total Population Estimates

Based on comparisons between statistics calculated from the NIS and the NHDS, it appears that most statistics calculated from the two data sources are similar. Overall, when compared with the NHDS, the NIS seems to estimate higher discharges for certain types of hospitals (government hospitals and large hospitals) and higher in-hospital mortality rates. The higher mortality estimates may be in part because the NIS tends to have higher estimates of discharges for "large" hospitals, and these patients may represent a somewhat different severity of illness than those in other hospitals.

Estimates of LOS and mortality by diagnosis and procedure groups show few significant differences. However, several estimates of discharges by diagnosis and procedure groups are significantly different. These differences of LOS and mortality could be attributable to differences in data handling — the NIS takes all diagnosis and procedure codes as they are recorded, while the NHDS has specific rules for what is considered a valid first-listed diagnosis.

Conclusion

In summary, the NIS estimates of ALOS appear to be unbiased in most contexts. The NIS estimates of discharge counts differ under some conditions from the NHDS estimates but not in any consistent direction. The NIS estimates for in-hospital mortality are higher than estimates from the NHDS for the Midwest and South. Based on comparisons with AHA data, NIS hospitals tend, on average, to be larger than the universe of community hospitals. This higher percentage of weighted NIS discharges coming from "large" hospitals — and the more complex case mix of those hospitals — may contribute to the higher in-hospital mortality estimates when compared to the NHDS.

REFERENCES

1. Gesler, Wilbert M. and Thomas C. Ricketts. *Health in Rural North America.* New Brunswick: Rutgers University Press, 1992.

2. Elixhauser, A. and McCarthy, E. *Clinical Classifications for Health Policy Research, Version 2: Hospital Inpatient Statistics.* (AHCPR Publication No. 96-0017) Agency for Health Care Policy and Research, Healthcare Cost and Utilization Project (HCUP) Research Note 1. February, 1996.

APPENDIX

Estimates of Standard Error for NHDS Statistics

A variety of statistics were estimated based on these data: 1) total number of discharges, 2) in-hospital mortality, and 3) average length of stay (calculated as the difference between discharge and admission dates). The standard errors were calculated as follows.

Total Numbers of Discharges

From the NHDS documentation, constants a and b were obtained for 1995. The standard error for the estimate of total discharges is:

$$SE_{TD} = \left(aW_{TD}^{2} + bW_{TD}\right)^{1/2}$$

where W_{TD} is the weighted sum of total discharges (i.e., the estimate of total discharges).

This estimate of standard error is valid only if:

(1) estimated total discharges exceeds 366,657 or

(2) estimated total discharges exceeds 60,769 and estimated total days exceeds 283,338.

Percent Mortality

Let P be the estimated proportion of in-hospital deaths. The standard error of this proportion expressed as a percent is:

$$SE_{P} = 100\left(\frac{c\ P\ (1 - P)}{W}\right)^{1/2}$$

Where the constant c is given by NHDS documentation. This estimate of the standard error is valid only if:

(1) estimated total discharges exceeds 366,657 and the estimated number of deaths exceeds zero, or

(2) both estimated total discharges and estimated total deaths exceed 60,769.

Average Length of Stay

Let ALOS be the estimated average length of stay based on a weighted number of discharges equal to TD. If the weighted sum of patient length of stay is TLOS, and

$$ALOS = \frac{TLOS}{TD}$$

then the estimated standard error is:

$$SE_{ALOS} = ALOS \left[\left(a_1 + \frac{b_1}{TD} \right) + \left(a_2 + \frac{b_2}{TLOS} \right) \right]^{1/2} .$$

Constants a_1, a_2, b_1, and b_2 were obtained from the NHDS documentation concerning standard error calculations for average length of stay.

Tests of Statistical Significance

To test for a statistically significant difference between an NIS estimate, X, and an NHDS estimate, Y, the following procedure was used. The difference is significant if

$$absolute\ value \left(\frac{X - Y}{\sqrt{SE_X{}^2 + SE_Y{}^2}} \right) \geq S$$

where SE_X is the estimated standard error for the NIS estimate and SE_Y is the estimated standard error of the NHDS estimate. S is equal to 1.96 for significance at the .05 level and S is equal to 2.576 for significance at the .01 level.

If a valid estimate of either standard error, SE_X or SE_Y, could not be obtained, then a significance test was not performed.

Table 58. Differences Between NIS – Release 4 and NHDS Files Used in This Analysis

CHARACTERISTIC	DATABASE	
	NIS – Release 4	**NHDS**
Intended Universe	Discharges from community hospitals as defined by the AHA - nonfederal, short-term general, or other special hospitals that are not a hospital unit of an institution.	Discharges from short-stay hospitals (hospitals with an average length of stay of less than 30 days), general-specialty (medical or surgical) hospitals, or children's hospitals. The NHDS does not include federal, military, and Veterans Administration hospitals, nor does it include hospital units of institutions (i.e., prison hospitals).
- Specialty hospitals and units	AHA community hospitals may be specialty hospitals. Some AHA community hospitals include specialty units - obstetrics/ gynecology; short-term rehabilitation; and ear, nose, and throat.	Includes discharges from a few specialty hospitals (i.e., psychiatric, maternity, alcohol/chemical dependency, orthopedic, and head injury rehabilitation hospitals).
- HMO enrollees	Included	Included
- Bedsize	No restriction on bedsize.	Must have at least six beds staffed for patient use.
Sample or Universe	Sample	Sample
Sampling Frame	19 states	50 states and the District of Columbia
Sample Design	By geographic region, control/ownership, location, teaching status, and bedsize (bedsize categories are specific to the hospital's location and teaching status). 938 hospitals.	Includes all hospitals with at least 1,000 beds or more than 40,000 discharges annually - plus an additional sample of hospitals based on a stratified three-stage design. Approximately 490 hospitals.
Discharges included in database	All discharges from sampled hospitals: approximately 6.7 million.	A sample of discharges from sampled hospitals: approximately 260,000 discharges.
Charges	Reported charges missing for some HMO enrollees.	Not reported

CHARACTERISTIC	DATABASE	
	NIS – Release 4	**NHDS**
Reassignment of diagnosis codes	None	Myocardial infarctions are reassigned to the principal diagnosis when other circulatory diagnoses are present. For women discharged after a delivery, a code of V27 (Outcome of Delivery) from the supplemental classification is entered as the second-listed code, with a code designating normal or abnormal delivery in the first-listed position. If the first-listed diagnosis was a symptom, it was reassigned as a secondary diagnosis.

Table 59. NIS and NHDS Comparisons by Region, 1995

	Number of Discharges in Thousands (Standard Error)		Average Length of Stay in Days (Standard Error)		In-Hospital Mortality Rate: Percent (Standard Error)	
	NIS	NHDS	NIS	NHDS	NIS	NHDS
U.S.	34,792 (461)	34,101 (1,092)	5.28 (.05)	5.05 (.27)	2.58 ** (.03)	2.39 (.06)
Census Region						
Midwest	7,492 (226)	7,743 (603)	6.39 * (.13)	4.92 (.63)	2.90 ** (.06)	2.34 (.10)
Northeast	8,296 (201)	7,689 (423)	5.07 (.06)	5.94 (.52)	2.41 (.04)	2.59 (.09)
South	12,260 (290)	12,542 (629)	5.12 (.05)	5.01 (.40)	2.74 ** (.04)	2.44 (.10)
West	6,344 (191)	6,128 (442)	4.60 (.20)	4.21 (.50)	2.13 (.08)	2.10 (.12)

*	Difference is significant at the 0.05 level.

**	Difference is significant at the 0.01 level.

Table 60. NIS and NHDS Comparisons by Hospital Characteristics, 1995

	Number of Discharges in Thousands (Standard Error)		Average Length of Stay in Days (Standard Error)		In-Hospital Mortality Rate: Percent (Standard Error)	
	NIS	NHDS	NIS	NHDS	NIS	NHDS
Control/Bedsize						
Private/ Investor-owned						
Total	3,689	3,631	4.94	5.04	2.70	2.53
	(107)	(124)	(.14)	(.28)	(.12)	(.18)
6 - 99 beds	634 **	831	4.91	4.53	2.49	2.80
	(33)	(34)	(.21)	(.29)	(.12)	(.30)
100 - 199 beds	1,544 **	1,172	4.80	5.06	2.78 *	2.41
	(51)	(45)	(.12)	(.30)	(.10)	(.52)
200 - 299 beds	1,055	893	4.90	5.48	2.53	2.18
	(108)	(36)	(.41)	(.34)	(.36)	(.64)
300 - 499 beds	381 **	735	5.34	5.04	3.00	2.56
	(129)	(31)	(.28)	(.32)	(.14)	(.35)
500+ beds	75 [a]	-	6.50 [a]	-	3.38 [a]	-
	(60)	(b)	(0.0)	(b)	(0.0)	(b)
Private/Nonprofit						
Total	26,091	26,132	5.25	5.06	2.58 **	2.38
	(436)	(839)	(.05)	(.27)	(.03)	(.07)
6 - 99 beds	2,483 **	4,324	4.41	4.73	2.74 **	2.09
	(92)	(146)	(.11)	(.26)	(.07)	(.16)
100 - 199 beds	5,039 **	6,301	5.08	4.65	2.57 *	2.21
	(184)	(209)	(.10)	(.25)	(.06)	(.14)
200 - 299 beds	5,091	5,281	5.18	5.06	2.55	2.52
	(340)	(176)	(.10)	(.27)	(.07)	(.16)
300 - 499 beds	8,026	7,184	5.12	5.30	2.50	2.53
	(425)	(237)	(.08)	(.28)	(.05)	(.14)
500+ beds	5,452 **	3,042	6.06	5.76	2.65	2.50
	(383)	(105)	(.15)	(.32)	(.07)	(.21)

	Number of Discharges in Thousands (Standard Error)		Average Length of Stay in Days (Standard Error)		In-Hospital Mortality Rate: Percent (Standard Error)	
	NIS	NHDS	NIS	NHDS	NIS	NHDS
Government/ Nonfederal						
Total	5,011 ** (199)	4,338 (146)	5.70 (.24)	5.04 (.27)	2.54 (.05)	2.53 (.18)
6 - 99 beds	1,320 ** (52)	1,645 (60)	5.71 (.86)	3.98 (.23)	2.62 (.09)	2.80 (.30)
100 - 199 beds	919 ** (61)	470 (22)	4.93 * (.15)	4.15 (.30)	2.43 (.10)	2.41 (.52)
200 - 299 beds	425 (95)	286 (15)	4.31 (.27)	4.96 (.39)	1.96 (.20)	2.18 (.64)
300 - 499 beds	872 * (88)	1,118 (43)	5.98 (.22)	5.93 (.35)	2.60 (.14)	2.56 (.35)
500+ beds	1,477 ** (186)	818 (34)	6.42 (.19)	6.49 (.40)	2.66 (.12)	2.13 (.37)

[a] A significance test was not performed because a valid standard error was not available.

[b] The NHDS sample size was too small to calculate a valid estimate of standard error.

* Difference is significant at the 0.05 level.

** Difference is significant at the 0.01 level.

Table 61. NIS and NHDS Comparisons by Patient Characteristics, 1995

	Number of Discharges in Thousands (Standard Error)		Average Length of Stay in Days (Standard Error)		In-Hospital Mortality Rate: Percent (Standard Error)	
	NIS	NHDS	NIS	NHDS	NIS	NHDS
Primary Payer						
Medicare	12,500	11,547	7.15	6.91	5.10	4.81
	(188)	(698)	(.09)	(.50)	(.04)	(.28)
Medicaid	6,452	5,588	4.80	4.53	1.09	.98
	(205)	(1186)	(.12)	(1.02)	(.04)	(.11)
Private Insurance	12,618	11,486	3.86	3.80	1.13	1.14
	(265)	(574)	(.03)	(.31)	(.02)	(.08)
Self-pay	1,799 *	1,281	4.44	4.38	1.62	1.50
	(109)	(193)	(.14)	(.93)	(.04)	(.22)
No charge	50 **	809	5.00	4.57	1.32	1.88
	(9)	(122)	(.48)	(.98)	(.19)	(.30)
Other	1,180 **	2,786	4.70	4.12	1.46 **	1.07
	(63)	(418)	(.11)	(.87)	(.08)	(.12)
Missing	194 *	604	5.30	4.73	1.05	1.44
	(38)	(172)	(.94)	(1.67)	(.15)	(.20)
Age Group						
Under 15 years	5,853	5,995	3.49	3.42	.41 **	.49
	(162)	(801)	(.08)	(1.05)	(.02)	(.02)
15 - 44 years	10,439	10,513	3.88	3.81	.61 **	.52
	(185)	(1,028)	(.06)	(.70)	(.02)	(.02)
45 - 64 years	5,915	6,108	5.67	5.52	2.27	2.28
	(88)	(695)	(.05)	(1.14)	(.03)	(.04)
65 years and over	12,584	11,484	7.09	6.80	5.38 **	5.15
	(188)	(1,231)	(.09)	(1.18)	(.04)	(.03)
Gender						
Male	14,441	13,970	5.68	5.42	3.13 *	2.94
	(185)	(936)	(.08)	(.53)	(.03)	(.07)
Female	20,345	20,131	5.00	4.80	2.20 *	2.01
	(292)	(640)	(.04)	(.26)	(.03)	(.07)
Race						
White	20,549	21,848	5.44	5.11	2.86 **	2.56
	(489)	(1,066)	(.06)	(.40)	(.03)	(.10)
Black	4,169	4,313	5.98	5.58	2.36	2.23
	(186)	(327)	(.12)	(.68)	(.04)	(.12)

	Number of Discharges in Thousands (Standard Error)		Average Length of Stay in Days (Standard Error)		In-Hospital Mortality Rate: Percent (Standard Error)	
	NIS	NHDS	NIS	NHDS	NIS	NHDS
Other	3,426 **	1,328	4.43	4.78	1.64 *	2.16
	(234)	(227)	(.13)	(1.34)	(.09)	(.21)
Missing	6,648	6,612	4.80	4.58	2.35 **	1.99
	(404)	(855)	(.07)	(.92)	(.05)	(.11)

* Difference is significant at the 0.05 level.

** Difference is significant at the 0.01 level.

Table 62. NIS and NHDS Comparisons by Hospital Characteristics and Patient Characteristics for South Region, 1995

	Number of Discharges in Thousands (Standard Error)		Average Length of Stay in Days (Standard Error)		In-Hospital Mortality Rate: Percent (Standard Error)	
	NIS	NHDS	NIS	NHDS	NIS	NHDS
Control/Ownership						
Private/	2,526	2,522	4.89	4.96	2.89	2.51
Investor-owned	(81)	(88)	(.09)	(.28)	(.07)	(.23)
Private/Nonprofit	7,435	7,362	4.98	4.94	2.68	2.42
	(231)	(242)	(.06)	(.28)	(.05)	(.13)
Government/	2,699	2,658	5.71	5.23	2.76	2.41
Nonfederal	(161)	(93)	(.16)	(.29)	(.08)	(.22)
Bedsize						
6 - 99 beds	1,715 **	2,390	4.89	4.31	2.92	2.75
	(52)	(84)	(.20)	(.24)	(.07)	(.25)
100 - 199 beds	2,863 *	3,156	4.84	4.69	2.81 **	2.09
	(80)	(109)	(.07)	(.26)	(.06)	(.19)
200 - 299 beds	1,936 *	2,347	4.89	5.06	2.66	2.41
	(173)	(83)	(.12)	(.29)	(.09)	(.23)
300 - 499 beds	3,128	2,681	5.04	5.37	2.73	2.55
	(252)	(93)	(.07)	(.30)	(.07)	(.23)
500+ beds	3,020 **	1,968	5.73	5.80	2.64	2.48
	(330)	(71)	(.14)	(.33)	(.10)	(.26)
Primary Payer						
Medicare	4,778	4,485	6.65	6.78	5.12	4.82
	(112)	(281)	(.05)	(.52)	(.05)	(.45)
Medicaid	2,202	2,191	4.78	4.11	1.13	.92
	(75)	(466)	(.17)	(.93)	(.03)	(.16)
Private Insurance	4,359	4,088	3.80	3.84	1.24	1.11
	(139)	(209)	(.05)	(.32)	(.04)	(.13)
Self-pay	902 **	172	4.42	4.35	1.72	1.81
	(102)	(27)	(.18)	(.98)	(.06)	(.65)
No charge	1 **	357	5.14	4.61	.88	1.53
	(0)	(55)	(1.09)	(1.01)	(.47)	(.41)
Other	407 **	958	4.68	4.11	1.99 **	1.01
	(20)	(145)	(.10)	(.88)	(.13)	(.21)
Missing	10 **	292	3.88	4.51	1.40	1.95
	(2)	(83)	(.31)	(1.59)	(.19)	(.34)

	Number of Discharges in Thousands (Standard Error)		Average Length of Stay in Days (Standard Error)		In-Hospital Mortality Rate: Percent (Standard Error)	
	NIS	NHDS	NIS	NHDS	NIS	NHDS
Age Group						
Under 15 years	2,042 (85)	2,096 (280)	4.03 (.18)	3.42 (1.05)	.45 (.02)	.51 (.04)
15 - 44 years	3,630 (116)	3,924 (384)	3.81 (.07)	3.70 (.68)	.75 ** (.04)	.58 (.04)
45 - 64 years	2,262 (56)	2,355 (268)	5.39 (.06)	5.49 (1.13)	2.35 (.03)	2.23 (.06)
65 years and over	4,726 (111)	4,167 (447)	6.46 (.05)	6.75 (1.17)	5.45 * (.06)	5.27 (.06)
Gender						
Male	5,316 (124)	5,068 (340)	5.42 (.06)	5.32 (.53)	3.34 ** (.04)	3.01 (.11)
Female	7,341 (173)	7,474 (242)	4.90 (.06)	4.79 (.26)	2.30 * (.04)	2.05 (.11)

*   Difference is significant at the 0.05 level.
** Difference is significant at the 0.01 level.

Table 63. NIS and NHDS Comparisons by Primary Diagnoses Ranked by NIS Data, 1995

Rank[1]	CCHPR Category[2]	Number of Discharges in Thousands (Standard Error)		Average Length of Stay in Days (Standard Error)		In-Hospital Mortality Rate: Percent (Standard Error)	
		NIS	NHDS	NIS	NHDS	NIS	NHDS
1	218: Liveborn	3,843 (114)	3,621 (123)	2.78 (.05)	2.78 (.16)	.34 (.01)	.38 (.08)
2	101: Coronary atherosclerosis	1,387 (45)	1,303 (49)	4.72 (.11)	4.42 (.26)	.97 (.02)	.95 (.20)
3	122: Pneumonia (except that caused by tuberculosis and sexually transmitted diseases)	1,268 (18)	1,261 (48)	6.89 (.06)	6.75 (.40)	6.39 (.10)	5.97 (.49)
4	108: Congestive heart failure, nonhypertensive	942 (15)	880 (36)	6.95 (.31)	6.35 (.39)	5.62 (.06)	4.69 (.53)
5	100: Acute myocardial infarction	720 (17)	766 (32)	6.52 (.06)	6.61 (.41)	9.49 (.11)	9.93 (.80)
6	193: Trauma to perineum and vulva	657 ** (20)	2 (1)	1.56[a] (.02)	1.47 (b)	0.0[a] (0.0)	0.0 (b)
7	109: Acute cerebrovascular disease	622 * (9)	554 (25)	8.58 (.14)	8.38 (.55)	11.28 (.14)	11.65 (1.01)
8	196: Normal pregnancy and/or delivery	611 ** (21)	3,763 (128)	1.53 ** (.01)	2.12 (.12)	0.0 (0.0)	.02 (.02)
9	69: Affective disorders	557 (23)	621 (27)	10.38 (.26)	9.79 (.62)	.10 (.02)	.06 (.07)
10	106: Cardiac dysrhythmias	554 (10)	559 (25)	4.14 (.05)	4.10 (.28)	1.26 (.03)	1.11 (.33)
11	127: Chronic obstructive pulmonary disease and bronchiectasis	516 (8)	553 (25`)	6.43 (.10)	6.15 (.41)	3.18 (.07)	3.28 (.56)
12	205: Spondylosis and back problems	507 (15)	515 (23)	3.74 (.06)	3.78 (.27)	.18 (.01)	.28 (.17)
13	102: Nonspecific chest pain	501 ** (11)	73 (7)	2.16[a] (.02)	1.54 (b)	.08[a] (.01)	.53 (b)
14	149: Biliary tract disease	494 (8)	509 (23)	4.86 (.04)	4.33 (.30)	.81 (.03)	.61 (.26)
15	55: Fluid and electrolyte disorders	481 ** (10)	571 (25)	4.87 (.06)	4.71 (.32)	3.41 (.10)	3.52 (.57)

Rank[1]	CCHPR Category[2]	Number of Discharges in Thousands (Standard Error)		Average Length of Stay in Days (Standard Error)		In-Hospital Mortality Rate: Percent (Standard Error)	
		NIS	NHDS	NIS	NHDS	NIS	NHDS
16	237: Complication of device, implant or graft	459 (14)	412 (20)	6.56 (.10)	6.21 (.44)	1.99 (.04)	1.25 (.40)
17	128: Asthma	443 * (13)	506 (23)	3.84 (.04)	3.71 (.26)	.43 (.02)	.23 (.16)
18	190: Fetal distress and abnormal forces of labor	422 ** (20)	4 (2)	2.36[a] (.04)	1.37 (b)	0.0[a] (0.0)	0.0 (b)
19	50: Diabetes mellitus with complications	410 (9)	407 (20)	6.68 (.16)	6.59 (.46)	1.71 (.05)	2.23 (.54)
20	159: Urinary tract infections	400 * (7)	444 (21)	5.50	5.50 (.39)	1.84 (.05)	2.59 (.56)
21	203: Osteoarthritis	385 (11)	354 (18)	5.85 (.10)	5.98 (.44)	.24[a] (.01)	.05 (b)
22	2: Septicemia (except labor)	378 ** (7)	308 (16)	8.81 (.09)	8.69 (.64)	14.07 (.17)	14.81 (1.50)
23	195: Other complications of birth, puerperium affecting management of the mother	370 ** (12)	52 (6)	2.12[a] (.04)	2.52 (b)	.03[a] (.01)	0.0 (b)
24	181: Other complications of pregnancy	352 ** (12)	161 (11)	2.32 (.04)	2.68 (.29)	.03[a] (.01)	0.0 (b)
25	45: Maintenance chemotherapy, radiotherapy	291 ** (14)	112 (9)	3.83 (.08)	3.89 (.44)	.71 (.04)	.60 (.54)

[1] NIS rank is based on number of discharges.

[2] Diagnoses classified according to *Clinical Classifications for Health Policy Research, Version 2* (see Elixhauser and McCarthy, 1996)

[a] A significance test was not performed because a valid standard error was not available.

[b] The NHDS sample size was too small to calculate a valid estimate of standard error.

* Difference is significant at the 0.05 level.

** Difference is significant at the 0.01 level.

Table 64. NIS and NHDS Comparisons by Primary Procedures Ranked by NIS Data, 1995

Rank [1]	CCHPR Category[2]	Number of Discharges in Thousands (Standard Error)		Average Length of Stay in Days (Standard Error)		In-Hospital Mortality Rate: Percent (Standard Error	
		NIS	NHDS	NIS	NHDS	NIS	NHDS
1	115: Circumcision	1,060 (31)	1,150 (44)	2.21 (.02)	2.17 (.14)	.01 [a] (0.0)	0.0 (b)
2	137: Other procedures to assist delivery	1,040 (45)	933 (37)	1.70 (.02)	1.75 (.12)	.01 (0.0)	.02 (.03)
3	134: Cesarean section	800 (40)	769 (32)	3.61 (.02)	3.60 (.24)	.02 (0.0)	.05 (.06)
4	133: Episiotomy	781 ** (35)	483 (22)	1.68 (.02)	1.62 (.13)	0.0 [a] (0.0)	0.0 (b)
5	47: Diagnostic cardiac catheterization, coronary arteriography	628 ** (22)	523 (24)	4.35 (.02)	3.95 (.28)	1.14 (.03)	.78 (.28)
6	70: Upper gastrointestinal endoscopy, biopsy	608 ** (9)	530 (24)	6.51 ** (.07)	5.30 (.36)	2.46 (.04)	1.81 (.43)
7	124: Hysterectomy, abdominal and vaginal	545 (13)	557 (25)	3.37 (.02)	3.39 (.24)	.12 (.01)	.11 (.10)
8	140: Repair of current obstetric laceration	512 * (20)	591 (26)	1.62 (.02)	1.65 (.13)	0.0 [a] (0.0)	0.0 (b)
9	45: Percutaneous transluminal coronary angioplasty (PTCA)	460 * (29)	383 (19)	4.31 (.05)	4.33 (.32)	1.11 (.05)	.93 (.36)
10	216: Respiratory intubation and mechanical ventilation	442 ** (9)	278 (15)	11.59 ** (.23)	8.56 (.65)	31.45 * (.41)	27.40 (1.98)
11	84: Cholecystectomy and common duct exploration	413 (8)	419 (20)	5.32 (.06)	4.73 (.34)	.84 (.03)	.42 (.23)
12	219: Alcohol and drug rehabilitation/detoxification	407 (31)	361 (18)	6.46 (.24)	7.16 (.51)	.09 [a] (.01)	0.0 (b)
13	231: Other therapeutic procedures	404 (40)	411 (20)	5.89 (.12)	5.53 (.39)	2.67 (.18)	2.63 (.58)
14	135: Forceps, vacuum, and breech delivery	393 (15)	398 (19)	1.84 (.02)	1.96 (.17)	.01 (0.0)	.01 (.04)
15	3: Laminectomy, excision intervertebral disc	357 (13)	318 (17)	3.58 (.06)	3.53 (.06)	.20 (.01)	.21 (.19)
16	44: Coronary artery bypass graft (CABG)	353 (21)	329 (17)	10.07 (.14)	9.58 (.69)	3.21 (.08)	2.20 (.60)

Rank [1]	CCHPR Category[2]	Number of Discharges in Thousands (Standard Error)		Average Length of Stay in Days (Standard Error)		In-Hospital Mortality Rate: Percent (Standard Error	
		NIS	NHDS	NIS	NHDS	NIS	NHDS
17	177: Computerized axial tomography (CT) scan head	314 * (19)	267 (15)	6.17 (.16)	5.70 (.45)	4.59 (.13)	4.11 (.90)
18	152: Arthroplasty knee	293 (8)	275 (15)	5.10 (.05)	4.97 (.40)	.21 [a] (.01)	.08 (b)
19	224: Cancer chemotherapy	285 * (14)	238 (14)	4.36 (.10)	3.98 (.34)	1.27 * (.06)	.47 (.33)
20	4: Diagnostic spinal tap	279 (9)	255 (14)	6.15 ** (.08)	5.01 (.41)	2.28 (.08)	1.57 (.58)
21	153: Hip replacement, total and partial	279 (7)	254 (14)	6.92 (.08)	7.11 (.56)	1.50 (.05)	1.61 (.58)
22	146: Treatment, fracture or dislocation of hip and femur	273 (4)	253 (14)	7.68 (.08)	7.56 (.59)	2.25 (.06)	2.89 (.78)
23	193: Diagnostic ultrasound of heart (echocardiogram)	247 (15)	236 (14)	5.71 (.10)	5.62 (.46)	2.27 (.09)	1.60 (.60)
24	76: Colonoscopy & biopsy	239 (4)	223 (13)	6.78 (.06)	6.15 (.51)	1.63 (.05)	.89 (.47)
25	217: Other respiratory therapy	192 ** (18)	264 (15)	4.78 (.14)	4.22 (.35)	3.47 * (.22)	1.86 (.61)

[1] NIS rank is based on number of discharges.

[2] Diagnoses classified according to *Clinical Classifications for Health Policy Research, Version 2* (see Elixhauser and McCarthy, 1996)

[a] A significance test was not performed because a valid standard error was not available.

[b] The NHDS sample size was too small to calculate a valid estimate of standard error.

* Difference is significant at the 0.05 level.

** Difference is significant at the 0.01 level.

Table 65. Number of Hospitals in NIS Frame and AHA Universe by Hospital Characteristics, 1995

	1995 AHA Universe	1995 Frame[1] Weighted	1995 Frame[1] Unweighted
U.S.	5,260	5,260	938
Census Region			
Midwest	1,507	1,507	479
Northeast	772	772	162
South	2,004	2,004	278
West	977	977	181
Control/Ownership			
Private/ investor-owned	785	772	145
Private/nonprofit	3,112	3,163	587
Government/ nonfederal	1,363	1,325	206
Location/Teaching Status			
Rural			
Total	2,257	2,257	367
1 - 49 beds	1,276	1,276	201
50 - 99 beds	570	570	97
100+ beds	411	411	69
Urban			
Total	3,003	3,003	571
Teaching			
Total	647	647	129
1 - 49 beds	258	258	50
50 - 99 beds	224	224	46
100+ beds	165	165	33
Non-teaching			
Total	2,356	2,356	442
1 - 49 beds	822	822	142
50 - 99 beds	780	780	160
100+ beds	754	754	140

Note: Significance tests were not performed because these are not sample statistics.

[1] The 1995 frame contains 19 states.

Table 66. NIS 19-State Sampling Frame and AHA Universe Comparisons, 1995

	Universe Mean	Frame Weighted Mean	Universe Median	Frame Weighted Median
Hospital Admissions	5852.29	6946.03	3250.00	4448.00
Hospital Discharges	5852.29	6946.03	3250.00	4448.00
Hospital Discharges[1]	6644.03	7887.94	3657.00	4986.00
Hospital Beds	151.73	175.47	96.00	122.0
Hospital Average Length of Stay	6.24	5.93	5.06	5.13
Hospital Occupancy	0.50	0.54	0.51	0.55
Total Hospital Expenses (in dollars)	54,145,873	66,091,226	24,687,389	34,682,636
Hospital Expenses per Bed (in dollars)	298,128	336,030	272,915	309,801
Total Hospital Payroll (in dollars)	23,418,937	28,558,285	10,322,839	15,011,000
Hospital Payroll per Bed (in dollars)	126,631	142,412	115,515	129,805
% Medicare Days	53.39	52.84	53.78	53.26
% Medicare Discharges	45.03	44.31	45.11	44.17
% Medicare Discharges[1]	40.71	40.04	40.17	39.57
% Medicaid Days	14.16	13.41	11.98	11.25
% Medicaid Discharges	15.94	14.95	14.67	13.72
% Medicaid Discharges[1]	14.13	13.18	13.02	12.19
FTE[2]	711.76	845.62	363.50	469.50
FTE[2]/Bed	4.26	4.54	3.98	4.20

Note: Significance tests were not performed because these are not sample statistics.

[1] Adjusted for well newborns.

[2] Full-time equivalents.